Revision History

The following table shows the revision history for this document.

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/26/2015</td>
<td>2015.1</td>
<td>Initial Xilinx release.</td>
</tr>
</tbody>
</table>
Table of Contents

Chapter 1: Introduction

Overview ................................................................. 7
Guide Organization .................................................... 8
Chapter 2: Platform and Memory Model ................................... 10
Overview .................................................................. 10
OpenCL Platform Model ............................................... 10
OpenCL Devices and FPGAs .......................................... 12
PCle Reference Device ................................................ 12
OpenCL Memory Model .............................................. 13
Host Memory .......................................................... 14
Global Memory ....................................................... 14
Constant Global Memory .......................................... 15
Local Memory ........................................................ 15
Private Memory ....................................................... 15

Chapter 2: Platform and Memory Model

Overview .................................................................. 10
OpenCL Platform Model ............................................... 10
OpenCL Devices and FPGAs .......................................... 12
PCle Reference Device ................................................ 12
OpenCL Memory Model .............................................. 13
Host Memory .......................................................... 14
Global Memory ....................................................... 14
Constant Global Memory .......................................... 15
Local Memory ........................................................ 15
Private Memory ....................................................... 15

Chapter 3: Compilation Flow

Overview .................................................................. 16
Starting SDaccel ....................................................... 16
Creating a Solution .................................................... 18
Adding Host Code ..................................................... 18
Creating a Kernel ..................................................... 19
Expressing a Kernel in OpenCL C ............................... 20
Expressing a Kernel in C/C++ ..................................... 20
Adding Kernel Code to a Solution .............................. 22
Creating the Xilinx OpenCL Compute Unit Binary Container .................................................... 22
Choosing the Kernel Execution Target .......................... 23
### Chapter 4: Application Emulation

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>27</td>
</tr>
<tr>
<td>CPU Emulation</td>
<td>27</td>
</tr>
<tr>
<td>Compile for CPU Emulation</td>
<td>28</td>
</tr>
<tr>
<td>Run CPU Emulation</td>
<td>28</td>
</tr>
<tr>
<td>Hardware Emulation</td>
<td>29</td>
</tr>
<tr>
<td>Compile for Hardware Emulation</td>
<td>29</td>
</tr>
<tr>
<td>Run Hardware Emulation</td>
<td>29</td>
</tr>
</tbody>
</table>

### Chapter 5: Performance Estimate

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>31</td>
</tr>
<tr>
<td>Generating the System Performance Estimate Report</td>
<td>31</td>
</tr>
<tr>
<td>Analyzing the Performance Estimate Report</td>
<td>31</td>
</tr>
<tr>
<td>Design and Target Device Summary</td>
<td>32</td>
</tr>
<tr>
<td>Kernel Summary</td>
<td>33</td>
</tr>
<tr>
<td>Timing Information</td>
<td>33</td>
</tr>
<tr>
<td>Latency Information</td>
<td>34</td>
</tr>
<tr>
<td>Area Information</td>
<td>34</td>
</tr>
</tbody>
</table>

### Chapter 6: Kernel Optimization

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>36</td>
</tr>
<tr>
<td>Work Group Size</td>
<td>36</td>
</tr>
<tr>
<td>Barriers</td>
<td>38</td>
</tr>
<tr>
<td>Loop Unrolling</td>
<td>39</td>
</tr>
<tr>
<td>Loop Pipelining</td>
<td>41</td>
</tr>
<tr>
<td>Work Item Pipelining</td>
<td>44</td>
</tr>
</tbody>
</table>

### Chapter 7: Memory Architecture Optimization

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>46</td>
</tr>
<tr>
<td>Multiple Memory Ports per Kernel</td>
<td>46</td>
</tr>
<tr>
<td>Adjustable Bitwidth for Memory Ports in a Kernel</td>
<td>48</td>
</tr>
<tr>
<td>On-Chip Global Memories</td>
<td>49</td>
</tr>
<tr>
<td>Partitioning Memories Inside of Compute Units</td>
<td>50</td>
</tr>
<tr>
<td>Cyclic Partitioning</td>
<td>51</td>
</tr>
<tr>
<td>Block Partitioning</td>
<td>52</td>
</tr>
<tr>
<td>Complete Partitioning</td>
<td>53</td>
</tr>
</tbody>
</table>

### Chapter 8: Application Debug in SDAccel

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>55</td>
</tr>
</tbody>
</table>
Chapter 9: Host Application Profiling in SDAccel

- API Calls Summary ............................................. 61
- Kernel Execution Summary .................................... 62
- Data Transfer Summary ........................................ 62
- Top Ten Kernel Execution Summary .......................... 62
- Top Ten Buffer Writes ......................................... 62
- Top Ten Buffer Reads ......................................... 62

Chapter 10: Xilinx OpenCL Compiler (xocc)

Appendix A: Command Reference

- Overview ............................................................. 66
- Solution Creation Commands ................................... 66
  create_solution .................................................. 66
- Add Target Device ................................................ 67
- Host Code Commands ............................................ 67
  add_files .......................................................... 67
- Host Code Properties .......................................... 68
  file_type .......................................................... 68
  host_cflags ....................................................... 68
- Kernel Definition Commands .................................. 69
  create_kernel .................................................... 69
- Kernel Code Commands ........................................ 69
  add_files .......................................................... 69
- Kernel Properties ................................................ 70
  kernel_flags ....................................................... 70
  max_memory_ports ............................................... 70
  memory_data_port_width ..................................... 70
- Binary Container Definition Commands ...................... 71
  create_opencl_binary .......................................... 71
- Binary Container Properties .................................. 71
  region .............................................................. 71
- Compute Unit Definition Commands .......................... 72
  create_compute_unit .......................................... 72
- Emulation Commands ........................................... 73
  compile_emulation ............................................. 73
  run_emulation .................................................. 73
- Reporting Commands .......................................... 74
  report_estimate ................................................ 74
System Compilation Commands .................................................. 75
  build_system ......................................................................... 75
  package_system .................................................................... 75
System Execution Commands ...................................................... 76
  run_system ........................................................................... 76

Appendix B: Devices
  SDAccel Supported Devices .................................................... 77

Appendix C: Additional Resources and Legal Notices
  Xilinx Resources ................................................................. 78
  Solution Centers ................................................................. 78
  References ........................................................................... 78
  Please Read: Important Legal Notices .................................... 78
Chaper 1

Introduction

Overview

Software is at the foundation of application specification and development. Whether the end application is targeted towards entertainment, gaming, or medicine, most products available today began as a software model or prototype that needed to be accelerated and executed on a hardware device. From this starting point, the software engineer is tasked with determining the execution device to get a solution to market and to achieve the highest possible degree of acceleration possible.

One traditional approach to accomplish this task has been to rely on processor clock frequency scaling. On its own, this approach has entered a state of diminishing returns, which has in turn led to the development of multi-core and heterogeneous computing devices. These architectures provide the software engineer with the possibility to more effectively trade-off performance and power for different form factors and computational loads. The one challenge in using these new computing architectures is the programming model of each device. At a fundamental level, all multi-core and heterogeneous computing devices require that the programmer rethink the problem to be solved in terms of explicit parallelism.

Recognizing the programming challenge of multi-core and heterogeneous compute devices, the Khronos Group industry consortium has developed the OpenCL™ programming standard. The OpenCL specification for multi-core and heterogeneous compute devices defines a single consistent programming model and system level abstraction for all hardware devices that support the standard. For a software engineer this means a single programming model to learn what can be directly used on devices from multiple vendors.

As specified by the OpenCL standard, any code that complies with the OpenCL specification is functionally portable and will execute on any computing device that supports the standard. Therefore, any code change is for performance optimization. The degree to which an OpenCL program needs to be modified for performance depends on the quality of the starting source code and the execution environment for the application.

Xilinx® is an active member of the Khronos Group, collaborating on the specification of OpenCL, and supports the compilation of OpenCL programs for Xilinx FPGA devices.
SDAccel™ is the Xilinx development environment for compiling OpenCL programs to execute on a Xilinx FPGA device.

Guide Organization

There are some differences between compiling a program for execution in an FPGA and a CPU/GPU environment. The following chapters in this guide describe how to use the SDAccel development environment to compile an OpenCL program for a Xilinx FPGA device.

This book is intended to document the features and usages of the Xilinx SDAccel Development Environment. It is assumed that the user already has a working knowledge of OpenCL. Though it includes some high level OpenCL concepts, it is not intended as an exhaustive technical guide on OpenCL. For more information on OpenCL, see the OpenCL specification available from the Khronos Group, and the OpenCL introductory videos available on the Xilinx website.

Chapter 2: Platform and Memory Model

Chapter 2, Platform and Memory Model introduces the OpenCL platform and memory model and how it applies to programs being executed on an FPGA. This chapter provides a brief introduction to the key concepts and how these concepts are mapped in SDAccel.

Chapter 3: Compilation Flow

Chapter 3, Compilation Flow introduces the SDAccel OpenCL development environment. This chapter explains the structure of an SDAccel solution and the steps to compile an OpenCL program for execution on a Xilinx FPGA device.

Chapter 4: Application Emulation

One of the key differences between compiling for a CPU/GPU and an FPGA is the lack of fixed computation architectures in the FPGA. SDAccel generates an application specific compute architecture that allows the programmer to fine tune an application not only for performance but also for power. Chapter 4, Application Emulation focuses on how a user can develop, emulate, and deploy both the software and the compute architecture generated for a specific application.

Chapter 5: Performance Estimate

While the OpenCL standard guarantees functional portability, the achieved performance of an application depends on the target device and on any device specific optimizations.
Chapter 5, Performance Estimate introduces the system performance estimation characteristics of SDAccel.

Chapter 6: Kernel Optimization

Performance optimization is essential to generating the best possible implementation with SDAccel. Chapter 6, Kernel Optimization describes the kernel optimization techniques available in SDAccel. Each technique is presented with an example that highlights the benefits of the optimization.

Chapter 7: Memory Architecture Optimization

SDAccel enables the developer to explore the impact of different memory architectures on kernel and application performance. Chapter 7, Memory Architecture Optimization describes how the memory architecture created by SDAccel can be modified to improve application performance. Techniques for improving the kernel access to memory as well as the implementation of memories in the FPGA fabric are described here.

Chapter 8: Application Debug in SDAccel

SDAccel supports host program debugging in all flows and kernel debugging in the CPU emulation flow with integrated GDB in command line mode Chapter 8, Application Debug in SDAccel describes the steps needed to setup application debug in SDAccel.

Chapter 9: Host Application Profiling in SDAccel

SDAccel runtime automatically collects profiling data on host applications. After host program finishes execution. Chapter 10, Xilinx OpenCL Compiler (xocc) describes how to access this data.

Appendix A: Command Reference

SDAccel is a command line based development environment for implementing applications expressed in the OpenCL programming paradigm on Xilinx FPGA devices. Appendix A, Command Reference provides a description of every command available in SDAccel.

Appendix B: Devices

SDAccel solutions are compiled against a target device. A device is the combination of board and infrastructure components on which the kernels of an application will be executed. Appendix B, Devices provides a list of devices available in the current version of SDAccel.
Chapter 2

Platform and Memory Model

Overview

The OpenCL™ standard describes all hardware compute resources capable of executing OpenCL applications using a common abstraction for defining a platform and the memory hierarchy. The platform is a logical abstraction model for the hardware executing the OpenCL application code. This model, which is common to all vendors implementing this standard, provides the application programmer with a unified view from which to analyze and understand how an application is mapped into hardware. Understanding how these concepts translate into physical implementations on the FPGA is necessary for application optimization.

This chapter provides a review of the OpenCL platform model and its extensions to FPGA devices. It explains the mapping of the OpenCL platform and memory model into an SDAccel™ generated implementation.

OpenCL Platform Model

The OpenCL platform model defines the logical representation of all hardware capable of executing an OpenCL program. At the most fundamental level all platforms are defined by the grouping of a processor and one or more devices. The host processor, which runs the OS for the system, is also responsible for the general bookkeeping and task launch duties associated with the execution of parallel programs such as OpenCL applications. The device is the element in the system on which the kernels of the application are executed. The device is further divided into a set of compute units. The number of compute units depends on the target hardware for a specific application. A compute unit is defined as the element in the hardware device onto which a work group of a kernel is executed. This device is responsible for executing the operations of the assigned work group to completion. In accordance to the OpenCL standard division of work groups into work items, a compute unit is further subdivided into processing elements. A processing element is the data path in the compute unit, which is responsible for executing the operations of one work item. A conceptual view of this model is shown in Figure 2-1.
Chapter 2: Platform and Memory Model

An OpenCL platform always starts with a host processor. For the case of platforms created with Xilinx® devices, the host processor is an x86 based processor communicating to the devices using PCIe®. The host processor has the following responsibilities:

- Manage the operating system and enable drivers for all devices.
- Execute the application host program.
- Set up all global memory buffers and manage data transfer between the host and the device.
- Monitor the status of all compute units in the system.

In all OpenCL platforms, the host processor tasks are executed using a common set of API functions. The implementation of the OpenCL API functions is provided by the hardware vendor and is referred to as the runtime library.

The OpenCL runtime library, which is provided by the hardware vendor, is the logical layer in a platform that is responsible for translating user commands described by the OpenCL API into hardware specific commands for a given device. For example, when the application programmer allocates a memory buffer using the clCreateBuffer API call, it is the responsibility of the runtime library to keep track of where the allocated buffer physically resides in the system, and of the mechanism required for buffer access. It is important for the application programmer to keep in mind that the OpenCL API is portable across vendors, but the runtime library provided by a vendor is not. Therefore, OpenCL applications have to be linked at compile time with the runtime library that is paired with the target execution device.

The other component of a platform is the device. A device in the context of OpenCL is the physical collection of hardware resources onto which the application kernels are executed.

Figure 2-1: OpenCL Platform Model
Chapter 2: Platform and Memory Model

A platform must have at least one device available for the execution of kernels. Also, per the OpenCL platform model, all devices in a platform do not have to be of identical type.

OpenCL Devices and FPGAs

In the context of CPU and GPU devices, the attributes of a device are fixed and the programmer has very little influence on what the device looks like. On the other hand, this characteristic of CPU/GPU systems makes it relatively easy to obtain an off-the-shelf board. The one major limitation of this style of device is that there is no direct connection between system I/O and the OpenCL kernels. All transactions of data are through memory based transfers.

An OpenCL device for an FPGA is not limited by the constraints of a CPU/GPU device. By taking advantage of the fact that the FPGA starts off as a blank computational canvas, the user can decide the level of device customization that is appropriate to support a single application or a class of applications. In determining the level of customization in a device, the programmer needs to keep in mind that kernel compute units are not placed in isolation within the FPGA fabric.

FPGA devices capable of supporting OpenCL programs consist of the following:

- Connection to the host processor
- I/O peripherals
- Memory controllers
- Interconnect
- Kernel region

The creation of FPGA devices requires FPGA design knowledge and is beyond the scope of capabilities for SDAccel. Devices for SDAccel are created using the Xilinx Vivado® design suite for FPGA designers. SDAccel provides pre-defined devices as well as allows users to augment the tool with third party created devices. A methodology guide describing how to create a device for SDAccel is available upon request from Xilinx.

The devices available in SDAccel are for Virtex®-7, and Kintex®-7 devices from Xilinx. These devices are available in a PCIe form factor. The PCIe form factor for Virtex-7 and Kintex-7 devices assumes that the host processor is an x86 based processor and that the FPGA is used for the implementation of compute units.

PCIe Reference Device

The PCIe base device has a distributed memory architecture, which is also found in GPU accelerated compute devices. This means that the host and the kernels access data from separate physical memory domains. Therefore, the developer has to be aware that passing
buffers between the host and a device triggers memory data copies between the physical memories of the host and the device. The data transfer time must be accounted for when determining the best optimization strategy for a given application. A representative example of this type of device is shown in Figure 2-2.

The main characteristics of devices with a PCIe form factor are as follows:

- The x86 processor in the PC is the host processor for the OpenCL application.
- The infrastructure IP provided as part of the device is needed for communication to the host over PCIe and to access the DDR memories on the board.
- Connecting OpenCL kernels to IP other than infrastructure IP or blocks generated by SDAccel is not supported.
- Kernels work on data in the DDR memory attached to the FPGA device.

OpenCL Memory Model

OpenCL defines the memory model to be used by all applications that comply with the standard. This hierarchical representation of memory is common across all vendors and can be applied to any OpenCL application. It is up to the vendor to define how the OpenCL memory model maps to specific hardware. The OpenCL memory model is shown in Figure 2-3 overlaid onto the OpenCL device model.
The memory hierarchy defined in OpenCL has the following levels:

- Host Memory
- Global Memory
- Constant Memory
- Local Memory
- Private Memory

**Host Memory**

The host memory is defined as the region of system memory that is only visible and accessible to the host processor. The host processor has full control of this memory space and can read and write from this space without any restrictions. Kernels cannot access data located in this space. Any data needed by a kernel must be transferred into global memory so that it is accessible by a compute unit.

**Global Memory**

The global memory is defined as the region of system memory that is accessible to both the OpenCL host and device. The host is responsible for the allocation and deallocation of buffers in this memory space. There is a handshake between host and device over control of the data stored in this memory. The host processor transfers data from the host memory space into the global memory space. Then, once a kernel is launched to process the data, the host loses access rights to the buffer in global memory. The device takes over and is
capable of reading and writing from the global memory until the kernel execution is complete. Upon completion of the operations associated with a kernel, the device turns control of the global memory buffer back to the host processor. Once it has regained control of a buffer, the host processor can read and write data to the buffer, transfer data back to the host memory, and deallocate the buffer.

**Constant Global Memory**

Constant global memory is defined as the region of system memory that is accessible with read and write access for the OpenCL host and with read only access for the OpenCL device. As the name implies, the typical use for this memory is to transfer constant data needed by kernel computation from the host to the device.

**Local Memory**

Local memory is defined as the region of system memory that is only accessible to the OpenCL device. The host processor has no visibility and no control on the operations that occur in this memory space. This memory space allows read and write operations by all the compute units within a device. This level of memory is typically used to store and transfer data that must be shared by multiple compute units.

**Private Memory**

Private memory is the region of system memory that is only accessible by a processing element within an OpenCL device. This memory space can be read from and written to by all of the work items scheduled for execution on the same processing element.

For devices using an FPGA device, the physical mapping of the OpenCL memory model is the following:

- Host memory is any memory connected to the host processor only.
- Global and constant memories are any memory that is connected to the FPGA device. These are usually memory chips that are physically connected to the FPGA device. The host processor has access to these memory banks through infrastructure in the FPGA base device.
- Local memory is memory inside of the FPGA device. This memory is typically implemented using block RAM elements in the FPGA fabric.
- Private memory is memory inside of the FPGA device. This memory is typically implemented using registers in the FPGA fabric in order to minimize latency to the compute data path in the processing element.
Chapter 3

Compilation Flow

Overview

SDAccel™ is the Xilinx® development environment for creating and compiling OpenCL™ applications onto a Xilinx FPGA. This tool suite provides a software development environment for algorithm development and emulation on x86 based workstations, as well as deployment mechanisms for Xilinx FPGA devices.

The compilation of OpenCL applications into binaries for execution on an FPGA does not assume nor require FPGA design knowledge. A basic understanding of the capabilities of an FPGA is necessary during application optimization in order to maximize performance. SDAccel handles the low level details of program compilation and optimization during the generation of application specific compute units for an FPGA fabric. Therefore, using SDAccel to compile an OpenCL program does not place any additional requirements on the user beyond what is expected for compilation towards a CPU or GPU target.

This chapter explains how an OpenCL program is compiled for execution on an FPGA by SDAccel.

Starting SDAccel

SDAccel is a command line based tool suite for compiling OpenCL programs into a Xilinx FPGA device. This tool suite works in a batch mode in which the user invokes the tool with a command file as in the following command:

```
    sdaccel <command file>
```

The command file is a text file which specifies a sequence of commands for SDAccel to execute during the process of compiling an OpenCL application. An example command file for SDAccel is shown below.

```
# Define a solution name for the current run of SDAccel
create_solution -name mysolution

# Set the target device for the current solution
add_device "xilinx:adm-pcie-7v3:1ddr:1.0"
```
Chapter 3: Compilation Flow

```
# Set any special host compilation flags
set_property -name host_cflags -value "-g -Wall -D FPGA_DEVICE" -objects [current_solution]

# Add host code
add_files test-cl.c

# Create an OpenCL Kernel Binary Container
create_opencl_binary mmult1

# Select the execution region within the target device
set_property region "OCL_REGION_0" [get_opencl_binary mmult1]

# Create a kernel
create_kernel -type clc mmult
# Add code to the kernel
add_files -kernel [get_kernels mmult] mmult1.cl

# Create a compute unit
create_compute_unit -opencl_binary [get_opencl_binary mmult1] -kernel [get_kernels mmult] -name myinstance

# Compile application for CPU based emulation
compile_emulation

# Compile the application for hardware emulation
compile_emulation -flow hardware

# Run the CPU based emulation
run_emulation

# Run the hardware emulation
run_emulation -flow hardware

# Generate system performance estimates
report_estimate

# Compile the application to run on an FPGA
build_system

# Package the solution
package_system

# Run the application in hardware
run_system
```

The commands in the script above cover all of the main compilation tasks in SDAccel. Each command is explained in detail throughout this guide.

Creating a Solution

The first step in any SDAccel design is to create a solution. A solution is a directory structure and application representation used by SDAccel to compile a design. The directory
structure, which is common to all SDAccel designs, is shown in Figure 3-1.

![SDAccel Directory Structure](image)

**Figure 3-1: SDAccel Directory Structure**

The command to start a new application is as follows:

```
create_solution -name mysolution
```

The `create_solution` command creates the environment in which an OpenCL application can be developed and compiled. In addition to generating the solution object for SDAccel to compile the application, the purpose of the solution is to define the device on which the compiled application is executed. A solution can only have one target device. Compiling the same application for multiple devices requires the application programmer to run SDAccel with one solution definition per board. A target device is added to a solution by the following command

```
add_device "xilinx:adm-pcie-7v3:1ddr:1.0"
```

A complete list of devices available in SDAccel is available in Appendix A, Command Reference.

---

**Adding Host Code**

An OpenCL application is composed of a host code base and a kernel code base. The host code base is the user code that runs on a processor and utilizes the OpenCL runtime API to enqueue and launch kernels to complete a specific computational workload. SDAccel compiles host code onto an x86 processor for application development and emulation. For deployment, SDAccel compiles the host code for the host processor specified in the target device. SDAccel supports devices with x86 based host processors.

The host code for an application is specified by the following:

```
add_files test-cl.c
```

By default, the `add_files` command marks files to be compiled onto the host processor. The path to the source file can be relative to the current directory or absolute. Compilation
Chapter 3: Compilation Flow

of host code files happens in memory, which means SDAccel does not require the user to modify how their source code is stored in any way.

One thing to keep in mind is that the `add_files` command does not make any distinction between source and header files. For proper compilation of the user code, SDAccel must know which files are to be treated as sources and which are to be treated as header files. The distinction between file types is defined by a property on a specific file. The command sequence to specify a file as a header file is as follows:

```plaintext
add_files test-cl.h
set_property file_type "c header files" [get_files "test-cl.h"]
```

**IMPORTANT:** SDAccel needs to know about all header and source files that are unique to the application being developed. Missing header and source files will lead to compilation errors when generating the binaries for the host processor and the kernels. Standard header files do not have to be developer specified.

Creating a Kernel

The computational foundation of every OpenCL application is a kernel. A kernel defines the code that will be accelerated and executed in parallel to complete a given application. SDAccel supports solutions with N kernels. Each kernel can be targeted for execution on a processor or the FPGA fabric depending on the resources available in the device chosen to execute the application. In addition, a kernel can be specified in either the OpenCL C kernel language or in C/C++ that is suitable for the Vivado® High-Level Synthesis Tool.

A kernel is created in the context of an SDAccel solution by the following:

```plaintext
create_kernel -type clc mmult
```

The `create_kernel` command takes as arguments the name of the kernel and the language in which it is specified. The name of the kernel must match the name of the top kernel function in the source code for which the programmer wants SDAccel to create a custom compute unit. The `type` parameter is used to tell the compiler which language front end to use in processing the application code. The choice of kernel language depends on the preference of the application programmer. SDAccel compiles both OpenCL C and C/C++ for Vivado® HLS with the same compiler.

Expressing a Kernel in OpenCL C

SDAccel supports the OpenCL language constructs and built in functions from the OpenCL 1.0 embedded profile.

---

**Send Feedback**

**Send Feedback**
The following is an example of an OpenCL kernel for matrix multiplication that can be compiled with SDAccel.

```c
__kernel __attribute__ ((reqd_work_group_size(16,16,1)))
void mult(__global int* a, __global int* b, __global int* output)
{
    int r = get_local_id(0);
    int c = get_local_id(1);
    int rank = get_local_size(0);
    int running = 0;

    for(int index = 0; index < 16; index++){
        int aIndex = r*rank + index;
        int bIndex = index*rank + c;
        running += a[aIndex] * b[bIndex];
    }
    output[r*rank + c] = running;
    return;
}
```

### Expressing a Kernel in C/C++

The kernel for matrix multiplication can be expressed in C/C++ code that can be synthesized by the Vivado HLS tool. For kernels captured in this way, SDAccel supports all of the optimization techniques available in Vivado HLS. The only thing that the user has to keep in mind is that expressing kernels in this way requires compliance with a specific function signature style.

```c
void mmult(int *a, int *b, int *output)
{
    #pragma HLS INTERFACE m_axi port=a offset=slave bundle=gmem
    #pragma HLS INTERFACE m_axi port=b offset=slave bundle=gmem
    #pragma HLS INTERFACE m_axi port=output offset=slave bundle=gmem
    #pragma HLS INTERFACE s_axilite port=a bundle=control
    #pragma HLS INTERFACE s_axilite port=b bundle=control
    #pragma HLS INTERFACE s_axilite port=output bundle=control
    #pragma HLS INTERFACE s_axilite port=return bundle=control

    const int rank = 16;
    int running = 0;
    int bufa[256];
    int bufb[256];
    int bufc[256];

    memcpy(bufa, (int *) a, 256*4);
    memcpy(bufb, (int *) b, 256*4);

    for (unsigned int c=0;c<rank;c++){
        for (unsigned int r=0;r<rank;r++){
            running=0;
            for (int index=0; index<rank; index++) {
                #pragma HLS pipeline
                running += a[bufa[index]] * b[bufb[index]];
                bufc[index] = running;
            }
            output[r*rank + c] = running;
        }
    }
    return;
}
```
int aIndex = r*rank + index;
int bIndex = index*rank + c;
running += bufa[aIndex] * bufb[bIndex];
}
bufc[r*rank + c] = running;
}
}

memcpy((int *) output, bufc, 256*4);
return;
}

The preceding code example is the matrix multiplication kernel expressed in C/C++ for Vivado HLS. The first thing to notice about this code is the function signature.

```c
void mmult(int *a, int *b, int *output)
```

This function signature is almost identical to the signature of the kernel expressed in OpenCL C. It is important to keep in mind that by default, kernels captured in C/C++ for HLS do not have any inherent assumptions on the physical interfaces that will be used to transport the function parameter data. HLS uses pragmas embedded in the code to direct the compiler as to which physical interface to generate for a function port. For the function to be treated as a valid OpenCL kernel, the ports on the C/C++ function must be reflected on the memory and control interface pragmas for HLS.

The memory interface specification is

```c
#pragma HLS INTERFACE m_axi port=<variable name> offset=slave bundle=<interface name>
```

The control interface specification is

```c
#pragma HLS INTERFACE s_axilite port=<variable name> bundle=<interface name>
```

Detailed information on how these pragmas are used is available in the Vivado Design Suite User Guide: High level Synthesis (UG902) [Ref 1].

**IMPORTANT:** Global variables in HLS C/C++ kernels are not supported.

---

### Adding Kernel Code to a Solution

As in the case of host code, SDAccel requires that the user add all source and header files that are associated with a kernel defined in the solution. This task is accomplished by the `add_files` command with an additional kernel parameter.

The code for a kernel in the solution is specified by the following command.

```bash
add_files -kernel [get_kernels mmult] mmult1.cl
```
The `add_files` command has the same behavior as with host code. The only difference is that code added with a `-kernel` flag is associated with a specific kernel object in the current solution. The path to the source file can be relative to the current directory or absolute. Compilation of kernel code files happens in memory, which means SDAccel does not require the user to modify how their source code is stored in any way.

One thing to keep in mind is that the `add_files` command does not make any distinction between source and header files. For proper compilation of the user code, SDAccel must know which files are to be treated as sources and which are to be treated as header files. The distinction between file types is defined by a property on a specific file. The command sequence to specify a file as a header file is as follows:

```
add_files -kernel [get_kernels mmult] mmult1.h
set_property file_type "c header files" [get_files "mmult1.h"]
```

**IMPORTANT:** SDAccel needs to know about all user source files for correct compilation of a kernel. This also applies to header files that are unique to the kernel.

### Creating the Xilinx OpenCL Compute Unit Binary Container

The main difference between targeting an OpenCL application to a CPU/GPU and targeting an FPGA is the source of the compiled kernel. Both CPUs and GPUs have a fixed computing architecture onto which the kernel code is mapped by a compiler. Therefore, OpenCL programs targeted for either kind of device invokes just in time compilation of kernel source files from the host code. The API for invoking just in time kernel compilation is as follows:

```
clCreateProgramWithSource(…)
```

In contrast to a CPU or a GPU, an FPGA can be thought of as a blank computing canvas onto which a compiler generates an optimized computing architecture for each kernel in the system. This inherent flexibility of the FPGA allows the developer to explore different kernel optimizations and compute unit combinations that are beyond what is possible with a fixed architecture. The only drawback to this flexibility is that the generation of a kernel specific optimized compute architecture takes a longer time than what is acceptable for just in time compilation. The OpenCL standard addresses this fundamental difference between devices by allowing for an offline compilation flow. This allows the user to generate libraries of kernels that can be loaded and executed by the host program. The OpenCL API for supporting kernels generated in an offline compilation flow is as follows:

```
clCreateProgramWithBinary(…)
```

SDAccel uses an offline compilation flow to generate kernel binaries. In order to maximize efficiency in the host program and allow the simultaneous instantiation of kernels that
cooperate in the computation of a portion of an application, Xilinx has defined the Xilinx OpenCL Compute Unit Binary format .xclbin. The xclbin file is a binary library of kernel compute units that will be loaded together into an OpenCL context for a specific device. This format can hold either programming files for the FPGA fabric or shared libraries for the processor. It also contains library descriptive metadata, which is used by the Xilinx OpenCL runtime library during program execution.

The library metadata included in the xclbin file is automatically generated by SDAccel and does not require user intervention. This data is composed of compute unit descriptive information that is automatically generated during compute unit synthesis and used by the runtime to understand the contents of an xclbin file.

The xclbin file is created using the following:

```
create_opencl_binary -device <device name> <binary name>
```

An SDAccel solution can generate N binary files. The appropriate number of binary files depends on the application and the amount of hardware resource time multiplexing that is desired. SDAccel assumes that all compute units stored within a binary can be executed concurrently on a given device. The name given during the invocation of the create_opencl_binary command defines the name of the xclbin file to be used by the host program. For the matrix multiplication example, the compiled kernel library xclbin is the following:

```
mmult.xclbin
```

The create_opencl_binary command does not launch the generation of any compute unit in the SDAccel solution. This command sets up the SDAccel solution with the target kernel library containers that will be created later in the compilation flow.

---

**Choosing the Kernel Execution Target**

SDAccel has the ability to compile kernels into compute units for execution on devices in an OpenCL device. The valid devices supported by SDAccel are processors and the FPGA fabric. SDAccel does not make a decision on where a compute unit is to be implemented. This decision is left to the application programmer and is communicated to the tool as part of the binary definition with the create_opencl_binary command.

The first stage of defining where compute units will be implemented is to define the area of a device that is targeted by a given xclbin file. By default, SDAccel targets the first available processing region in a device for the implementation of compute units. For devices with more than one region for compute units, the user can select a specific region with the following command:

```
set_property region <region name> [get_opencl_binary <binary name>]
```
The name of a region in the property described above is provided as part of the device. Since SDAccel can handle the generation of multiple xclbin binary containers per solution, the user must annotate this property with the name of the binary file to which it applies.

The second stage of defining where a kernel will be executed is to instantiate compute units in the system. A compute unit is the custom hardware data path that SDAccel generates to implement the functionality in the kernel code. SDAccel supports the generation of up to N compute units for any kernel and there is no limit on the number of kernels and compute units that an application can have. The limitation on compute units which can be executed concurrently is placed by the device developer who defines how many of the available device resources can be dedicated to the implementation of compute units. SDAccel issues a warning if the number of compute units packed into a single binary container exceeds the computational resources available in the target device. A compute unit for a kernel is defined by the following command:

```
create_compute_unit -opencl_binary [get_opencl_binary <binary name>]
-kernel [get_kernels <kernel name>]
-name <instance name>
```

This command creates a compute unit for a specific kernel into the specified binary container. The `name` field for a compute unit instance name is optional. If the application programmer only wants one compute unit for a given kernel, then the `name` field can be omitted from the `create_compute_unit` command. For cases where more than one compute unit is required for a kernel, the `create_compute_unit` command can be repeated as many times as necessary. The only requirement is that the `name` field is used with a unique name every time the `create_compute_unit` command is used for the same kernel. Also, a compute unit instance name cannot be used more than once per binary container.

The specifics of how a kernel is compiled and synthesized into a custom compute unit depend on the region of the device selected during the definition of the binary container. The `create_compute_unit` command does not launch the compilation and synthesis of kernels into compute units and binary containers. This is the last solution setup command in SDAccel.

---

**Building the System**

Compilation of an application for execution on a Xilinx enabled OpenCL device through SDAccel is called the build system step. This step goes beyond compilation of host and kernel code and is also responsible for the generation of custom compute units for all of the binary containers in the solution.

The following is the command for this step in the flow:

```
build_system
```
For each binary container in the solution, the `build_system` command takes inventory of all of the compute units that must be generated and their execution region in the OpenCL device. The `build_system` command completes after all compute units and binary containers defined in a solution have been created.

The compilation method used for each kernel by the `build_system` command is dependent on the user selected kernel execution target. The `build_system` commands invokes different flows for kernels targeted at a processor and kernels targeted at the FPGA fabric.

### Build Flow for Compute Units Targeting the FPGA Fabric

SDAccel generates custom logic for every compute unit in the binary container. Therefore, it is normal for this build step to run for a longer period of time than the other steps in the SDAccel compilation flow.

The steps in compiling compute units targeting the FPGA fabric are as follows:

1. Generate a custom compute unit for a specific kernel.
2. Instantiate the compute units in the OpenCL binary container.
3. Connect the compute units to memory and infrastructure elements of the target device.
4. Generate the FPGA programming file.

The generation of custom compute units for any given kernel code utilizes the production proven capabilities of the Xilinx Vivado High Level Synthesis (HLS) tool, which is the compute unit generator in SDAccel. Based on the characteristics of the target device in the solution, SDAccel invokes the compute unit compiler to generate custom logic that maximizes performance while at the same time minimizing compute resource consumption on the FPGA fabric. Automatic optimization of a compute unit for maximum performance is not possible for all coding styles without additional user input to the compiler. Chapter 6, **Kernel Optimization** discusses the additional user input that can be provided to SDAccel to optimize the implementation of kernel operations into a custom compute unit.

Once all compute units have been generated, these units are connected to the infrastructure elements provided by the target device in the solution. The infrastructure elements in a device are all of the memory, control, and I/O data planes which the device developer has defined to support an OpenCL application. SDAccel combines the custom compute units and the base device infrastructure to generate an FPGA binary which is used to program the Xilinx device during application execution.

**IMPORTANT:** SDAccel always generates a valid FPGA hardware design, but does not generate an optimal allocation of the available bandwidth in the control and memory data planes. The user can manually optimize the data bandwidth utilization by selecting connection points into the memory and control data planes per compute unit.
Package System

The final stage of SDAccel application compilation is to package the generated system for execution and deployment on FPGA based accelerator cards to be run outside of the SDAccel environment. The command for this step is the following:

```
package_system
```

The output of the build_system command is stored in the following location:

```
<solution name>/pkg
```

All files necessary to execute the application on a Xilinx OpenCL device outside of the SDAccel environment are provided at this location.

Run System

Once the application has been compiled and packaged by SDAccel, the user has the choice of launching the application from within SDAccel or running it on a Linux console. To run the application from within SDAccel the command use the following:

```
run_system -args <command line arguments>
```

The `run_system` command executes the compiled user application with any command line arguments added by the `args` flag. This command is only valid for applications compiled against PCIe® based devices and requires the PCIe based accelerator card to be connected to the development workstation.
Overview

One of the key differences between compiling for a CPU/GPU and an FPGA is the lack of fixed computation architectures in the FPGA. SDAccel™ leverages the inherent flexibility of the FPGA to generate a kernel specific compute architecture that allows the programmer to fine tune an application not only for performance, but also for power. While there are many benefits to executing an application on an FPGA, this execution device does introduce a design challenge not found in CPU/GPU development.

The challenge presented by FPGA devices is that the development and execution of an application occur on two distinct systems. Application development happens on the application programmer’s development workstation, and deployment can occur on remote servers loaded with FPGA based accelerator cards. The difference in environment during an application’s life cycle introduces the need for emulation before the application is deployed.

This chapter describes how SDAccel enables emulation of an OpenCL™ application targeted to an FPGA. In the context of the compilation flow described in Chapter 3, Compilation Flow, the commands in this chapter are to be executed before the build_system command.

CPU Emulation

In the context of SDAccel, application emulation on a CPU is the same as the iterative development process that is typical of CPU/GPU programming. In this type of development style, a programmer continuously compiles and runs an application as it is being developed.

Although partitioning and optimizing an application into kernels is integral to OpenCL development, performance is not the main goal at this stage of application development in SDAccel. The main goal of CPU emulation is to ensure functional correctness and to partition the application into kernels.
For CPU based emulation, both the host code and the kernel code are compiled to run on an x86 processor. The programmer model of iterative algorithm refinement through fast compile and run loops is preserved with speeds that are the same as a CPU compile and run cycle. The steps to enable this development style in SDAccel are the following:

- Compile for emulation
- Run the emulation

### Compile for CPU Emulation

The CPU based emulation flow in SDAccel is only concerned with the functional correctness of an application. Therefore, for the purpose of checking functionality, all kernels in an application are compiled into compute units that are executed as threads on an x86 based processor. SDAccel leverages the functional portability of the OpenCL standard to abstract the details of compiling for emulation from the application programmer. The same host code that will be deployed when the application runs on the FPGA based accelerator card can be used for all of the emulation flows in SDAccel.

Compiling the solution for CPU based emulation is executed by the following command:

```
compile_emulation
```

Since this emulation is based on all parts of the application executing on an x86 processor, the compilation time is very fast and similar to what is expected when compiling a program for CPU / GPU execution.

### Run CPU Emulation

Once the application has been compiled using the `compile_emulation` command, it can be executed within the SDAccel environment to verify functional correctness. The command to execute the CPU based emulation in SDAccel is the following:

```
run_emulation -args <command line arguments>
```

The `run_emulation` command executes the CPU emulation of an OpenCL application. This command has an additional `args` parameter, which is optional. The purpose of the `args` parameters is to enable the application programmer to pass any command line options to the host program which would also be used during application deployment.

The `run_emulation` command runs the host code program to completion and invokes the Xilinx® OpenCL runtime library as well as all necessary kernels to complete the functionality. This command does not do any checks on correctness of the application execution. It is the responsibility of the application programmer to verify the results of the program.

---

**CAUTION!** SDAccel has no concept of what a user application does or what constitutes correct behavior. The programmer must check application results for correctness.
Chapter 4: Application Emulation

Hardware Emulation

SDAccel generates at least one custom compute unit for each kernel in an application. This means that while the CPU emulation flow is a good measure of functional correctness, it does not guarantee correctness on the FPGA execution target. Before deployment, the application programmer should also check that the custom compute units generated by the tool are producing the correct results.

SDAccel has a hardware emulation flow, which enables the programmer to check the correctness of the logic generated for the custom compute units. This emulation flow invokes the hardware simulator in SDAccel to test the functionality of the logic that will be executed on the FPGA compute fabric. The steps in the hardware emulation flow are the following:

- Compile for hardware emulation
- Run the hardware emulation

Compile for Hardware Emulation

The FPGA system created by SDAccel for a given application consists of device infrastructure and custom compute units packed into the Xilinx OpenCL Binary Container. The current hardware emulation flow is based on checking the correctness of the compute units generated by SDAccel. Correctness of the device is the responsibility of the device developer and is beyond the current scope of SDAccel.

Compiling the solution for hardware emulation is executed by the following command:

```
compile_emulation -flow hardware
```

Hardware emulation requires SDAccel to generate the custom logic for each compute unit in the system. It is expected for this step in the SDAccel flow to take several minutes.

Run Hardware Emulation

Once the application has been compiled for hardware emulation, it can be executed within the SDAccel environment to verify functional correctness of the custom compute units. The command to execute the CPU based emulation in SDAccel is the following:

```
run_emulation -flow hardware -args <command line arguments>
```

The `run_emulation` command with the hardware flow qualifier executes the hardware emulation of an OpenCL application. This command has an additional `args` parameter, which is optional. The purpose of the `args` parameters is to enable the application programmer to pass any command line options to the host program which would also be used during application deployment.
The run_emulation command will run the host code program to completion and invoke the Xilinx OpenCL runtime library as well as all necessary kernels to complete the functionality. It is important to keep in mind that the compute units implementing the kernel operations are executed within a hardware simulator for this type of emulation. Therefore, the time it takes to complete this command is directly proportional to the amount of data transferred from the host to the kernels and the number of operations in a compute unit.

The run_emulation command does not do any checks on correctness of the application execution. It is the responsibility of the application programmer to verify the results of the program.

**CAUTION!** SDAccel has no concept of what a user application does or what constitutes correct behavior. The programmer must check application results for correctness.
Performance Estimate

Overview

The generation of FPGA programming files is the step in SDAccel™ with the longest execution time. It is also the step in which the execution time is most affected by the target hardware device and the number of compute units placed on the FPGA fabric. Therefore, it is essential for the application programmer to have an understanding of the performance of the application before running it on the target device. This chapter introduces the profiling capabilities in SDAccel.

Generating the System Performance Estimate Report

The system performance estimate in SDAccel takes into account the target hardware device and each compute unit in the application. Although an exact performance metric can only be measured on the target device, the estimation report in SDAccel provides an accurate representation of the expected behavior. For the benefits of system level estimation before system implementation to be seen, the generation of the system performance estimate report precedes the `build_system` command in the compilation flow. The command to generate the system performance estimate report is:

```
report_estimate
```

The system performance estimate report is stored in:

```
<solution name>/rpt
```

Analyzing the Performance Estimate Report

The performance estimate report generated by the `report_estimate` command provides information on every binary container in the application, as well as every compute unit in the design. The structure of the report is:
Chapter 5: Performance Estimate

- Target device information
- Summary of every kernel in the application
- Detailed information on every binary container in the solution

The following example report file represents the information that is generated by the `report_estimate` command.

```
Design Name:     myproj
Target Device: vc690-admpcie7v3-1ddr-gen2
Target Board:    admpcie7v3
Target Clock:    167MHz
Solution Created: Mon Aug 18 14:57:00 PDT 2015

-------------------------------------------------------------
Kernel Summary
Total number of kernels: 1

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>Type</th>
<th>Target</th>
<th>Binary Container</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>clc</td>
<td>FPGA</td>
<td>test_container</td>
</tr>
</tbody>
</table>

-------------------------------------------------------------
OpenCL Binary = test_container
Kernels mapped to: clc_region
Timing Information (MHz)

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>Target Frequency</th>
<th>Estimated Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>167</td>
<td>190.84</td>
</tr>
</tbody>
</table>

Latency Information (in clock cycles)

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>Start Interval</th>
<th>Best Case</th>
<th>Avg Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>924</td>
<td>923</td>
<td>923</td>
</tr>
</tbody>
</table>

Area Information

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>FF</th>
<th>LUT</th>
<th>DSP</th>
<th>BRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>924</td>
<td>1244</td>
<td>923</td>
<td>923</td>
</tr>
</tbody>
</table>
```

Design and Target Device Summary

All design estimate reports begin with an application summary and information about the target hardware. Device information is provided by the following section of the report:

```
Design Name: myproj
Target Device: vc690-admpcie7v3-1ddr-gen2
Target Board: admpcie7v3
Target Clock: 167MHz
Solution Created: Mon Aug 18 14:57:00 PDT 2015
```
For the design summary, the only information you provide is the design name and the selection of the target device. The other information provided in this section is the target board and the clock frequency.

The target board is the name of the board which will run the application compiled by SDAccel. The clock frequency defines how fast the logic for compute units mapped to the FPGA fabric will run. Both of these parameters are fixed by the device developer. These parameters cannot be modified from within SDAccel.

**Kernel Summary**

The kernel summary section lists all of the kernels defined for the current SDAccel solution. An example kernel summary is shown below.

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>Type</th>
<th>Target</th>
<th>Binary Container</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>clc</td>
<td>FPGA</td>
<td>test_container</td>
</tr>
</tbody>
</table>

Along with the kernel name, the summary provides the execution target and the OpenCL™ binary container where the compute unit of the kernel is stored. Also, since there is a difference in compilation and optimization methodology for OpenCL C and C/C++ source files, the type of kernel source file is specified.

The kernel summary section is the last summary information in the report. From here, detailed information on each compute unit binary container is presented.

**Timing Information**

The detail section for each binary container begins with the execution target of all compute units. It also provides timing information for every compute unit. As a general rule, an estimated frequency that is higher than that of the device target means that the compute unit will run in hardware. If the estimated frequency is below the target frequency, the kernel code for the compute unit will need to be further optimized before the compute unit will run correctly on the FPGA fabric. An example of this information is

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>Target Frequency</th>
<th>Estimated Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>167</td>
<td>194</td>
</tr>
</tbody>
</table>

The importance of the timing information is the difference between the target and the estimated frequencies. As stated in Chapter 2, Platform and Memory Model, compute units are not placed in isolation into the FPGA fabric. Compute units are placed as part of a valid FPGA design that can include other components defined by the device developer to support a class of applications.
Since the compute unit custom logic is generated one kernel at a time, an estimated frequency that is higher than the device target provides the SDAccel developer with confidence that there will not be any problems during the creation of the FPGA programming files.

**Latency Information**

The latency information presents the execution profile of each compute unit in the binary container. When analyzing this data, it is important to keep in mind that all values are measured from the compute unit boundary through the custom logic. In-system latencies associated with data transfers to global memory are not reported as part of these values. Also, the latency numbers reported are only for compute units targeted at the FPGA fabric. An example of the latency report is:

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>Start Interval</th>
<th>Best Case</th>
<th>Avg Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>924</td>
<td>923</td>
<td>923</td>
</tr>
</tbody>
</table>

The latency report is divided into the following fields:

- Start internal
- Best case latency
- Average case latency

The start interval defines the amount of time that has to pass between invocations of a compute unit for a given kernel. This number sets the limit as to how fast the runtime can issue application ND range data tiles to a compute unit.

The best and average case latency numbers refer to how much time it takes the compute unit to generate the results of one ND range data tile for the kernel. For cases where the kernel does not have data dependent computation loops, the latency values will be the same. Data dependent execution of loops introduces data specific latency variation that is captured by the latency report.

**Area Information**

Although the FPGA can be thought of as a blank computational canvas, there are a limited number of fundamental building blocks available in each FPGA device. These fundamental blocks (FF, LUT, DSP, block RAM) are used by SDAccel to generate the custom logic for each compute unit in the design. The number of each fundamental resource needed to implement the custom logic in a compute unit determines how many compute units can be
simultaneously loaded into the FPGA fabric. An example of the area information reported for a compute unit is shown below:

Area Information

<table>
<thead>
<tr>
<th>Compute Unit</th>
<th>Kernel Name</th>
<th>FF</th>
<th>LUT</th>
<th>DSP</th>
<th>BRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>test</td>
<td>924</td>
<td>1244</td>
<td>923</td>
<td>923</td>
</tr>
</tbody>
</table>
Chapter 6

Kernel Optimization

Overview

The OpenCL™ standard guarantees functional portability but not performance portability. Therefore, even though the same code will run on every device supporting OpenCL, the performance achieved will vary depending on coding style and capabilities of the underlying hardware. Optimizing for an FPGA using the SDAccel™ tool chain requires the same effort as code optimization for a CPU/GPU. The one difference in optimization for these devices is that in a CPU/GPU, the programmer is trying to get the best mapping of an application onto a fixed architecture. For an FPGA, the programmer is concerned with guiding the compiler to generate optimized compute architecture for each kernel in the application.

This chapter introduces the methodology by which SDAccel handles some of the key constructs and attributes for the OpenCL kernel language. The kernel optimization techniques available in SDAccel are also presented.

Work Group Size

The workgroup size in the OpenCL standard defines the size of the ND range space that can be handled by a single invocation of a kernel compute unit. In CPU and GPU based devices, the attribute on the kernel to define the workgroup size is typically optional. The runtime sets the workgroup size based on device characteristics and launches parallel threads on the hardware to complete the application.

In the case of an FPGA implementation, the specification of the workgroup size is optional but highly recommended. The attribute is recommended for performance optimization during the generation of the custom logic for a kernel.
Starting with the following kernel code:

```c
__kernel __attribute__((reqd_work_group_size(4,4,1)))
void mmult32(__global int* A, __global int* B,
__global int* C)
{
    // 2D Thread ID
    int i = get_local_id(0);
    int j = get_local_id(1);
    __local int Blocal[256];
    int result=0, k=0;
    Blocal[i*16 + j] = B[i*16 + j];
    barrier(CLK_LOCAL_MEM_FENCE);
    for(k=0;k<16;k++) result += A[i*16+k]*B_local[k*16+j];
    C[i*16+j] = result;
}
```

SDAccel will generate custom logic to implement kernel functionality with or without the required workgroup size attribute. In order to generate an accelerator capable of handling an ND range problem space tile, SDAccel converts the kernel code into a form that is acceptable to the kernel compiler, which is based on the Vivado® HLS tool. The code transformation starts with the function signature and the workgroup size.

```c
__kernel __attribute__((reqd_work_group_size(4,4,1)))
void mmult32(global int *A,global int *B,global int *C)
{
    ...
}
```

This code is transformed into:

```c
__kernel void mmult32(global int* A,global int* B,global int* C)
{
    localid_t id;
    int B_local[16*16];
        for(id[1]=0;id[1]<4;id[1]++)
            for(id[0]=0;id[0]<4;id[0]++){
                ...
            }
    ...
}
```

The three for loops introduced by SDAccel into the kernel code are necessary to handle the three-dimensional space of the ND range. The programmer has to keep in mind that at runtime the compute unit in the FPGA fabric can only handle as many threads as the generated hardware can handle. Once the implementation is generated, there is no way to dynamically reallocate device resources at runtime.

The loop nest introduced by SDAccel can have either variable or fixed loop bounds. By setting the `reqd_work_group_size` attribute, the programmer is setting the loop boundaries on this loop nest. Fixed boundaries allow the kernel compiler to optimize the
size of local memory in the compute unit and to provide latency estimates. If the workgroup size is not specified, SDAccel will assume a large size for the local memory, which can hinder the number of compute units that can be instantiated in the FPGA fabric. In addition, the latency numbers for the kernel will not be calculated by the `report_estimate` command, because latency will be assumed to be data dependent.

---

**Barriers**

Barriers are the OpenCL kernel language constructs used to ensure memory consistency. When a kernel is launched, there is no guarantee on the order of execution for all work items in a kernel. Work items can be executed sequentially, in parallel, or in a combination of sequential and parallel execution. The execution start of a work item depends on the underlying hardware device. Since order of execution cannot be guaranteed across multiple devices, the OpenCL standard introduces the concept of barriers to create check points for memory consistency.

By definition, a barrier states that all work items must arrive at the barrier before any work item is allowed to proceed with the rest of the computation task. This ensures memory consistency and prevents any work item from overwriting data required by a different work item before it has been read. A barrier in OpenCL kernel code is expressed as:

```c
__kernel __attribute__ reqd_work_group_size(4,4,1)
void mmult32(global int *A, global int *B, global int *C)
{
    // 2D Thread ID
    int i = get_local_id(0);
    int j = get_local_id(1);
    __local int B_local[16*16];

    if(i == 0 && j == 0)
        async_work_group_copy(B_local,B,256, 0);

    barrier(CLK_LOCAL_MEM_FENCE);
    for(k=0;k<16;k++)
        result += A[i*16+k] * B_local[k*16+j];

    C[i*16+k] = result;
}
```
SDAccel transforms the barrier call into:

```c
__kernel void mmult32(global int* A, global int* B, global int* C)
{
    localid_t id;
    int B_local[16*16];

        for(id[1]=0;id[1]<4;id[1]++)
            for(id[0]=0;id[0]<4;id[0]++)
                { int i = get_local_id(id,0);
                  int j = get_local_id(id,1);
                  if(i == 0 && j == 0)
                      async_work_group_copy(B_local,B,256, 0);
                }

        for(id[1]=0;id[1]<4;id[1]++)
            for(id[0]=0;id[0]<4;id[0]++)
                { int i = get_local_id(id,0);
                  int j = get_local_id(id,1);
                  for(k=0;k<16;k++)
                      result += A[i*16+k] * B_local[k*16+j];
                  C[i*16+k] = result;
                }
}
```

The size of the work group determines the boundaries on both loop nests. SDAccel executes loop nests in sequential order, which preserves the memory consistency properties of a barrier.

---

**Loop Unrolling**

Loop unrolling is the first optimization technique available in SDAccel. The purpose of the loop unroll optimization is to expose concurrency to the compiler. This is an official attribute in the OpenCL 2.0 specification.

For example, starting with the code:

```c
/* vector multiply */
kerneld void
vmult(local int* a, local int* b, local int* c)
{
    int tid = get_global_id(0);
    for (int i=0; i<4; i++)
        { int idx = tid*4 + i;
          a[idx] = b[idx] * c[idx];
        }
}
```

The execution profile of the example is marked by four loop boundaries. This profile is similar to Figure 6-1.
In the execution profile shown in Figure 6-1, loop iterations execute on the same hardware elements in a purely sequential manner. Even if there is a possibility of running computations from multiple iterations in parallel, everything is executed sequentially because there is no concurrency being exposed to the compiler. This type of loop execution consumes the least amount of FPGA fabric resources. In this type of execution, not all of the hardware elements of the custom kernel logic are working all the time, which increases latency and decreases performance.

The execution profile of the loop can be improved by using the loop unroll attribute with an unroll factor:

```c
kernel void
vmult(local int* a, local int* b, local int* c)
{
    int tid = get_global_id(0);
    __attribute__((opencl_unroll_hint(2)))
    for (int i=0; i<4; i++) {
        int idx = tid*4 + i;
        a[idx] = b[idx] * c[idx];
    }
}
```

The code above tells SDAccel to unroll the loop by a factor of two. This results in two loop iterations instead of four for the compute unit to complete the operation. The execution profile of this code is as shown in Figure 6-2.
By enabling SDAccel to reduce the loop iteration count to two, the programmer has exposed more concurrency to the compiler. The newly exposed concurrency reduces latency, improves performance, and consumes more FPGA fabric resources.

The last hint that you can provide SDAccel with this attribute is to unroll the loop completely. The syntax for the fully unrolled version of the code example is as shown below:

```c
kernel void vmult(local int* a, local int* b, local int* c)
{
    int tid = get_global_id(0);
    __attribute__((opencl_unroll_hint))
    for (int i=0; i<4; i++) {
        int idx = tid*4 + i;
        a[idx] = b[idx] * c[idx];
    }
}
```

Fully unrolling the loop nest results in a single iteration with the execution profile shown in Figure 6-3.

![Figure 6-3: Loop Execution Profile after Complete Unrolling](image)

In this case, all of the possible concurrency in the loop nest is exposed to the compiler. SDAccel analyzes the data and control dependencies of the unrolled loop nest and automatically makes parallel all operations that can be executed concurrently.

---

**Loop Pipelining**

Although loop unrolling exposes concurrency, it does not address the issue of keeping all elements in a kernel logic busy at all times, which is necessary for maximizing application throughput. Even in an unrolled case, loop control dependencies can lead to sequential behavior. The sequential behavior of operations results in idle hardware and a loss of performance.
Chapter 6: Kernel Optimization

Xilinx addresses the issue of idle hardware resources by introducing a vendor extension on top of the OpenCL 2.0 specification for loop pipelining. The Xilinx attribute for loop pipelining is:

```
xcl_pipeline_loop
```

In order to understand the effect of loop pipelining on performance, consider the following code example:

```c
kernel void foo(...) {
  ...
  for (int i=0; i<3; i++) {
    int idx = get_global_id(0)*3 + i;
    op_Read(idx);
    op_Compute(idx);
    op_Write(idx);
  }
  ...
}
```

This kernel code has no attributes and is executed sequentially per the order of operations stated in the code fragment. The execution profile of this code is as shown in Figure 6-4.

Assuming each operation in the loop takes one clock cycle, one loop iteration takes three clock cycles, and the complete loop takes nine clock cycles to complete. Although the execution is functionally correct, the implementation is not maximizing performance because the read, compute, and write stages of the example are not always busy.
The pipeline attribute serves as a command to the SDAccel compiler to maximize performance and minimize the idle time of any stage in the generated logic. The example code with the loop pipeline attribute looks like:

```c
kernel void foo(...) {

  __attribute__((xcl_pipeline_loop))
  for (int i=0; i<3; i++) {
    int idx = get_global_id(0)*3 + i;
    op_Read(idx);
    op_Compute(idx);
    op_Write(idx);
  }

  ...}
```

The execution profile after loop pipelining is as shown in Figure 6-5.

![Execution Profile after Loop Pipelining](image)

**Figure 6-5: Execution Profile after Loop Pipelining**

From Figure 6-5, it can be seen that each stage of the loop body is kept busy all of the time in a pipelined fashion. While analyzing the impact of loop pipelining, it is important to keep in mind the two cycle numbers shown in Figure 6-5. The 1 clock cycle measurement is the amount of time that must pass between the start of loop iteration I and loop iteration I + 1. Without using unrolling to expose all possible concurrency, the best possible number that can be achieved for this metric is 1. The 5 clock cycle number is the overall latency to completely execute the loop nest. By adding loop pipelining to this example, the latency of the implementation is decreased by 45% without any code modification or additional hardware.
Work Item Pipelining

Work item pipelining is the extension of loop pipelining to the kernel work group. The syntax for the attribute for this optimization is:

\[ \text{xcl\_pipeline\_workitems} \]

An example of code where work pipelining can be applied is:

```c
kernel __attribute__((reqd_work_group_size(3,1,1)))
void foo(...) 
{...
  int tid = get_global_id(0);
  op_Read(tid);
  op_Compute(tid);
  op_Write(tid);
}...
}
```

In order to handle the `reqd_work_group_size` attribute, SDAccel automatically inserts a loop nest to handle the three-dimensional characteristics of the ND range. As a result of the loop nest added by SDAccel, the execution profile of this code is the same as that of an unpipelined loop (see Figure 6-6).

![Figure 6-6: Work Item Execution without Pipelining](image)

The work item pipeline attribute can be added to the code as:

```c
kernel __attribute__((reqd_work_group_size(3,1,1)))
void foo(...) 
{...
  __attribute__((xcl_pipeline_workitems)) {
    int tid = get_global_id(0);
    op_Read(tid);
    op_Compute(tid);
    op_Write(tid);
  }
  ...
}
```
As in the case of loop pipelining, the resulting hardware is kept busy all of the time, which maximizes performance, and results in the following execution profile, as shown in Figure 6-7.

**IMPORTANT:** Attributes are programmer hints to the OpenCL compiler for performance optimization. It is up to each individual compiler to follow or ignore the attribute. All attributes used by SDAccel, which are not part of the OpenCL standard start with the tag xcl.
Memory Architecture Optimization

Overview

Along with the compiler improvements available through the use of attributes, you can also increase the performance of an OpenCL™ application by modifying the memory bandwidth exposed from the system to the kernel. In the case of kernels mapped into the FPGA logic, SDAccel™ generates the custom hardware for the kernel computation and the connections to system-level memory. This key difference between SDAccel and OpenCL compilers for fixed architectures such as a GPU allows you to take full advantage of the parallel nature of the FPGA architecture and the power benefits derived from it.

There are three memory optimizations available in SDAccel:

- Multiple Memory Ports per Kernel
- Adjustable Bitwidth for Memory Ports in a Kernel
- On-Chip Global Memories

Multiple Memory Ports per Kernel

The default behavior of SDAccel is to generate functionally correct FPGA kernels that consume the least amount of FPGA resources. This behavior produces the most area efficient kernels, but might not always achieve the application throughput requirement. The area efficient behavior of SDAccel is demonstrated by the following example:

```c
__kernel
void example (__global float16 *A, __global float16 *B, __global float16 *C)
{
    int id = get_global_id(0);
    C[id] = A[id] * B[id];
}
```

The code above uses the float16 vector data type to encapsulate the multiplication of 16 values from buffer A by 16 values from buffer B. All 16 results are then stored into the appropriate locations in buffer C. The hardware implementation generated by SDAccel without additional help from the developer has a single port to global memory through
which all accesses to buffers A, B, and C are scheduled. Although float16 is a vector type, SDAccel uses the base type float to determine the size of the generated memory interface port. In this case, the memory port is generated with a 32-bit width. During execution, the width of the port forces sequential access to global memory to fetch individual elements from buffers A and B to compute the result C.

The first technique to improve the performance of this code is to use the `pipeline_work_items` attribute from Chapter 6, Work Item Pipelining. This compiler hint tells SDAccel that you want the multiplication operation in the work item to be scheduled as fast as possible to maximize use of the computation element. The application of work item pipelining is shown below:

```
__kernel
void example (__global float16 *A, __global float16 *B, __global float16 *C)
{
  __attribute__((xcl_pipeline_workitems)) {
    int id = get_global_id(0);
    C[id] = A[id] * B[id];
  }
}
```

Scheduling the multiplication operation to happen as quickly as possible requires simultaneous access to buffers A and B. The issue with this requirement is that the default behavior of SDAccel creates a single physical memory port for the kernel. The single physical port creates a bottleneck that forces sequential accesses to buffers A and B. SDAccel detects this system-level issue and alerts you with a warning message displayed on the console.

The message highlights that the throughput of the kernel is limited by the available memory bandwidth. SDAccel will not error out on this issue and only generates the warning message. In cases such as this, SDAccel automatically lowers the throughput of the application until an improved constraint set is implemented in hardware. This provides you with a valid working system that can be analyzed for performance bottlenecks.

One way of increasing the memory bandwidth available to a kernel is to increase the number of physical connections to memory that are attached to a kernel. Proper implementation of this optimization requires your knowledge of both the application and the target compute device. Therefore, SDAccel requires direct user intervention to increase the number of physical memory ports in a kernel. The SDAccel command to increase the number of physical memory ports available to the kernel is:

```
set_property max_memory_ports true [get_kernels <kernel name>]
```

The `max_memory_ports` property tells SDAccel to generate one physical memory interface for every global memory buffer declared in the kernel function signature. This command is only valid for kernels that have been placed into binaries that will be executed in the FPGA logic. There is no effect on kernels executing in a processor.
For the example in this chapter, the kernel function signature is:

```c
void example (__global float16 *A, __global float16 *B, __global float16 *C)
```

which contains global memory buffers A, B, and C. The SDAccel command sequence to generate a kernel with three physical memory ports is:

```
create_kernel -type clc example
set_property max_memory_ports true [get_kernels example]
```

The first command in the sequence creates a kernel example in the current solution. The property on the kernel guides SDAccel to create a physical memory port for each buffer in the kernel code. This property applies to all compute units created for the example kernel and that are placed into a binary targeted for the FPGA logic. Each compute unit for the example kernel will have parallel access to buffers A, B, and C as a result of this kernel property.

### Adjustable Bitwidth for Memory Ports in a Kernel

In addition to increasing the number of memory ports available to a kernel, you have the ability to change the bitwidth of the memory port. The benefit of modifying the bitwidth of the memory interface depends on the computation in the kernel. For this example, the core computation is:

```
```

where A, B and C are of type float16. A float16 is a 16 element vector data type created from the fundamental C data type float. SDAccel uses the fundamental C data type when determining the default bitwidth of a memory interface. In this case, the memory interfaces have a bitwidth of 32 bits. Therefore, the kernel requires 16 memory transactions to read enough data to complete the computation of a single work item. You can override the default behavior of SDAccel with the following kernel property command:

```
set_property memory_port_data_width <bitwidth> [get_kernels <kernel name>]
```

The bitwidths currently supported by SDAccel are 32, 64, 128, 256, and 512 bits. In cases where your defined bitwidth does not match the bitwidth declared in the kernel source code, SDAccel handles all data width conversions between the physical interface size and the data type in the kernel source code. This optimization is only supported for kernels mapped for execution in the FPGA logic.

The command sequence to enable the memory port bitwidth optimization is:

```
create_kernel -type clc example
set_property max_memory_ports true [get_kernels example]
set_property memory_port_data_width 512 [get_kernels example]
```
Chapter 7: Memory Architecture Optimization

The first command in the sequence creates a kernel example in the current solution. The property on the kernel guides SDAccel to create a physical memory port for each buffer in the kernel code. This property applies to all compute units created for the example kernel and that are placed into a binary targeted for the FPGA logic. Each compute unit for the example kernel will have parallel access to buffers A, B, and C as a result of this kernel property. In addition, the second property specified on the kernel sets the bitwidth of every physical memory port to 512 bits. As in the case of the max_memory_port property, changes to the memory_port_data_width property affect all compute units generated for a given kernel. The memory data width change reduces the number of memory transactions per buffer from 16 to 1 for every work item invocation.

IMPORTANT: Changes to the bitwidth of a kernel physical memory interface must always be specified after all memory ports have been generated. Kernels accessing more than one global memory buffer should set the max_memory_ports property before the bitwidth property is set.

On-Chip Global Memories

Another memory architectural optimization that is available in SDAccel deals with global memories that are used to pass data between kernels. In cases where the global memory buffer used for inter-kernel communication does not need to be visible to the host processor, SDAccel enables you to move the buffer out of DDR based memory and into the FPGA logic. This optimization is called on-chip global memory buffers and is part of the OpenCL 2.0 specification.

The on-chip global memory buffer optimization makes use of the block memory instances embedded in the FPGA logic to create a memory buffer that is only visible to the kernels accessing the buffer. The following code example illustrates the usage model for global memory buffers that is suitable for the on-chip global memory buffer optimization.

```c
// Global memory buffers used to transfer data between kernels
// Contents of the memory do not need to be accessed by host processor
global int g_var0[1024];
global int g_var1[1024];

// Kernel reads data from global memory buffer written by the host processor
// Kernel writes data into global buffer consumed by another kernel
kernel __attribute__ ((reqd_work_group_size(256,1,1)))
void input_stage (global int *input)
{
    __attribute__((xcl_pipeline_workitems)) {
    g_var0[get_local_id(0)] = input[get_local_id(0)];
    }
}

// Kernel computes a result based on data from the input_stage kernel
kernel __attribute__ ((reqd_work_group_size(256,1,1)))
void adder_stage(int inc)
{
    // code for adder stage
}
```
Chapter 7: Memory Architecture Optimization

__attribute__((xcl_pipeline_workitems)) {
    int input_data, output_data;
    input_data = g_var0[get_local_id(0)];
    output_data = input_data + inc;
    g_var1[get_local_id(0)] = output_data;
}

// Kernel writes the results computed by the adder_stage to a global memory buffer that is read by the host processor
kernel __attribute__((reqd_work_group_size(256,1,1)))
void output_state(global int *output) {
    __attribute__((xcl_pipeline_workitems)) {
        output[get_local_id(0)] = g_var1[get_local_id(0)];
    }
}

In the code example above, the input_stage kernel reads the contents of global memory buffer input and writes them into global memory buffer g_var0. The contents of buffer g_var0 are used in a computation by the adder_stage kernel and stored into buffer g_var1. The contents of g_var1 are then read by the output_stage kernel and stored into the output global memory buffer. Although both g_var0 and g_var1 are declared as global memories, the host processor only needs to have access to the input and output buffers. Therefore, for this application to run correctly the host processor must only be involved in setting up the input and output buffers in DDR based memory. Since buffers g_var0 and g_var1 are only used for inter-kernel communication, the accesses to these buffers can be removed from the system-level memory bandwidth requirements. SDAccel automatically analyzes this kind of coding style to infer that both g_var0 and g_var1 can be implemented as on-chip memory buffers. The only requirements on the memories are that all kernels with access to the on-chip memory are executed in the FPGA logic and that the memory has at least 4096 bytes. For example, in an array of int data type, the minimum array size needs to be 1024.

IMPORTANT: On-chip global memory optimization is an automatic optimization in SDAccel that is applied to memory buffers that are only accessed by kernels executed in the FPGA logic.

Partitioning Memories Inside of Compute Units

One of the advantages of the FPGA over other compute devices for OpenCL programs is the ability for the application programmer to customize the memory architecture all throughout the system and into the compute unit. By default, SDAccel generates a memory architecture within the compute unit that maximizes local and private memory bandwidth based on static code analysis of the kernel code. Further optimization of these memories is possible based on attributes in the kernel source code, which can be used to specify physical layouts and implementations of local and private memories. The attribute in
SDAccel to control the physical layout of memories in a compute unit is `array_partition`.

```
__attribute__((xcl_array_partition(<partition type>,
    <partition factor>,
    <array dimension>)))
```

The `array_partition` attribute implements an array declared within kernel code as multiple physical memories instead of a single physical memory. The selection of which partitioning scheme to use depends on the specific application and its performance goals. The array partitioning schemes available in SDAccel are cyclic, block, and complete.

### Cyclic Partitioning

Cyclic partitioning is the implementation of an array as a set of smaller physical memories that can be accessed simultaneously by the logic in the compute unit. The process of separating the original array into smaller physical memories is similar to the process of dealing a deck of cards among N players. The attribute for cyclic partitioning is

```
__attribute__((xcl_array_partition(cyclic,
    <partition factor>,
    <array dimension>)))
```

When the `array_partition` attribute is used in cyclic mode, the partition factor specifies among how many physical memories or players to split the original array in the kernel code. The array dimension parameter specifies on which dimension of the array to apply the portioning. SDAccel supports arrays of N dimensions and can modify the physical mapping of an array on any dimension.

For example, consider the following array declaration:

```
__local int buffer[16];
```

The array named `buffer` stores 16 values that are 32-bits wide each. Cyclic partitioning can be applied to this buffer as follows:

```
__local int buffer[16] __attribute__((xcl_array_partition(cyclic,4,1)));
```

The cyclic partitioning attribute above tells SDAccel to distribute the contents of `buffer` among four physical memories. The allocation of buffer elements, which follows the card dealing paradigm, is shown in Figure 7-1.
This attribute increases the immediate memory bandwidth for operations accessing the array buffer by a factor of four. All arrays inside of a compute unit in the context of SDAccel are capable of sustaining a maximum of two concurrent accesses. By dividing the original array in the code into four physical memories, the resulting compute unit can sustain a maximum of eight concurrent accesses to the array buffer.

**Block Partitioning**

Block partitioning is the physical implementation of an array as a set of smaller memories that can be accessed simultaneously by the logic inside of the compute unit. This type of partitioning is expressed in the kernel code by the following attribute:

```c
__attribute__((xcl_array_partition(block,
    <partition factor>,
    <array dimension>)))
```

When the `array_partition` attribute is used in block mode, the partition factor specifies the number of elements from the original array to store in each physical memory. The number of physical memories created by this optimization is given by the division of the original array size by the block size stated in the attribute. For example, consider the following array declaration:

```c
__local int buffer[16];
```

This is the same array as in the cyclic partitioning example. Apply block partitioning as follows:

```c
__local int buffer[16] __attribute__((xcl_array_partition(block,4,1)));
```

Since the size of the block is four, SDAccel will generate four physical memories with the data distribution shown in Figure 7-2.
For this example, both block and cyclic partitioning result in the same number of physical memories. Both schemes deliver a theoretical maximum of eight concurrent memory transactions. The choice of which to employ depends on the data access pattern of the compute unit. SDAccel will always build the partitioning scheme stated by the attribute, but does not guarantee maximum concurrency of memory operations. SDAccel schedules memory operations by taking into account both the physical layout of data and the access pattern described in the algorithm to ensure fidelity to the intent of the code and avoid data corruption issues. Therefore, depending on the functionality captured in the kernel, one partitioning scheme might deliver higher performance than the other at the same cost in terms of FPGA resources.

**Complete Partitioning**

Complete partitioning is the physical implementation of a buffer in the kernel code as independent registers within the custom logic of the compute unit. This type of partitioning is expressed in the kernel code by the following attribute:

```c
__attribute__((xcl_array_partition(complete, <array dimension>)))
```

When the array partition attribute is used in complete mode, the array in the kernel code is broken up into individual registers that are embedded into the data path of the compute unit. The number of individual registers created by SDAccel corresponds to the number of entries in the array declaration. For example, consider the following array declaration:

```c
__local int buffer[16];
```

This array, which is the same one used in the examples for block and cyclic partitioning, can be qualified for complete partitioning as follows:

```c
__local int buffer[16] __attribute__((xcl_array_partition(complete,1)));
```

The resulting memory layout into individual registers is shown in Figure 7-3.
For this example, complete partitioning results in 16 independent registers. While this creates an implementation with the highest possible memory bandwidth, it is not applicable to all applications. The way in which data is accessed by the kernel code through either constant or data dependent indexes affects the amount of supporting logic that SDAccel has to build around each register to ensure functional equivalence with the usage in the original code. As a general best practice guideline for SDAccel, the complete partitioning attribute is best suited for arrays in which at least one dimension of the array is accessed through the use of constant indexes.

![Physical Layout of Buffer After Complete Partitioning](image)

**Figure 7-3: Physical Layout of Buffer After Complete Partitioning**
Chapter 8

Application Debug in SDAccel

Overview

There are four steps to debugging applications in SDAccel™.

• Prepare Host Application for Debug
• Prepare Kernel Code for Debug in CPU Emulation Flow
• Launch GDB in SDAccel Environment to Debug Host and Kernel Programs
• Launch GDB Standalone to Debug Host Program with Kernel Running on FPGA

SDAccel supports host program debugging in all flows and kernel debugging in the CPU emulation flow with integrated GDB in command line mode.

Prepare Host Application for Debug

The host program needs to be compiled with debugging information generated in the executable by adding the \(-g\) option to the \(host\_cflags\) property in the SDAccel compilation script:

\[
\text{set\_property -name host\_cflags -value \{-g\} -objects \{current\_solution\}}
\]

Prepare Kernel Code for Debug in CPU Emulation Flow

Kernel code can be debugged together with the host program in the CPU emulation flow using integrated GDB. Debugging information needs to be generated first in the binary container by passing \(-g\) option to the \(kernel\_flags\) property on the kernel before it can be debugged.

\[
\text{set\_property -name kernel\_flags -value \{-g\} -objects \{get\_kernels smithwaterman\}}
\]
Launch GDB in SDAccel Environment to Debug Host and Kernel Programs

After host and kernel compilation are completed, GDB can be launched from the SDAccel environment by passing the -debug option to the run_emulation command. Below are examples for the CPU and hardware emulation flows respectively.

```bash
run_emulation -debug -flow cpu -args "test.xclbin"
run_emulation -debug -flow hardware -args "test.xclbin"
```

Figure 8-1 shows the GDB terminal window after the run_emulation command is executed. Please refer to GDB project web site https://www.gnu.org/software/gdb/ on GDB commands and usages.

![GDB Terminal Window](image)

The kernel function can be debugged like a regular C function inside GDB. Please note that GDB may ask "Make breakpoint pending on future shared library load?"
when you set a breakpoint on kernel functions. Answer “y” and press Enter. See the following example.

(gdb) break smithwaterman
Function "smithwaterman" not defined.
Make breakpoint pending on future shared library load? (y or [n]) y
Breakpoint 1 (smithwaterman) pending.
(gdb) r

Launch GDB Standalone to Debug Host Program with Kernel Running on FPGA

When running the SDAccel host program with the kernel running on an FPGA, the host program is generally packaged using the package_system command and often generated on a different machine. GBD can be launched standalone to debug the host program in this case if the host machine has the same version of SDAccel installed, and the host program was built with debug information (built with the -g flag). Below are the instructions:

• Set up your SDAccel environment by following instructions in SDAccel Development Environment Installation and Licensing Guide, (UG1020)[Ref 2].

• Launch GDB from the SDAccel installation on the host program as in the following example:

   <SDAccel_install_dir>/lnx64/tools/bin/gdb --args host.exe test.xclbin

Use printf() for Debugging Kernels

SDAccel™ supports OpenCL printf() kernel built-in function in all development flows: CPU emulation, hardware emulation and running kernel in actual hardware.

Below is a kernel example of using printf() and the output when the kernel is executed with global size of 8:

```c
__kernel __attribute__ ((reqd_work_group_size(1, 1, 1)))
void hello_world(__global int *a)
{
    int idx = get_global_id(0);

    printf("Hello world from work item %d\n", idx);
    a[idx] = idx;
}
```
Chapter 8: Application Debug in SDAccel

Output is as follows:

Hello world from work item 0
Hello world from work item 1
Hello world from work item 2
Hello world from work item 3
Hello world from work item 4
Hello world from work item 5
Hello world from work item 6
Hello world from work item 7
Host Application Profiling in SDAccel

The SDAccel™ runtime automatically collects profiling data on host applications. After the host application finishes execution, the profile summary is saved in both HTML and CSV formats in the solution report directory or working directory. This can be reviewed in a web browser or Microsoft Excel or OpenOffice.

- When running CPU emulation using the `run_emulation` command in SDAccel environment, the profile summary reports are as follows:
  
  `<solution_name>/rpt/profile_summary_cpu_em.html`
  `<solution_name>/rpt/profile_summary_cpu_em.csv`

- When running hardware emulation using the `run_emulation` command in the SDAccel environment, the profile summary reports are as follows:

  `<solution_name>/rpt/profile_summary_hw_emu.html`
  `<solution_name>/rpt/profile_summary_hw_emu.csv`

- When running an application on hardware using the `run_system` command in the SDAccel environment, the profile summary reports are as follows:

  `<solution_name>/rpt/profile_summary_hw.html`
  `<solution_name>/rpt/profile_summary_hw.csv`

- When running an application standalone in all flows outside the SDAccel environment, the following profile summary reports are generated in the directory where the application is executed:

  `<working_directory>/sdaccel_profile_summary.html`
  `<working_directory>/sdaccel_profile_summary.csv`
Chapter 9: Host Application Profiling in SDAccel

The following report shows the SDAccel Profile Summary in a web browser:

SDAccel Profile Summary

Generated on: 2015-04-07 11:14:06

Profiled application:
/wrk/work1/smith_waterman/baseline_project_gdb/impl/slm/alg/baseline_project_gdb.exe

API Calls Summary

<table>
<thead>
<tr>
<th>API Name</th>
<th>Number Of Calls</th>
<th>Total Time (ms)</th>
<th>Average Time (ms)</th>
<th>Maximum Time (ms)</th>
<th>Minimum Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>clCreateProgramWithBinary</td>
<td>1</td>
<td>46.215</td>
<td>46.215</td>
<td>46.215</td>
<td>46.215</td>
</tr>
<tr>
<td>clWaitForEvents</td>
<td>1</td>
<td>0.798</td>
<td>0.798</td>
<td>0.798</td>
<td>0.798</td>
</tr>
<tr>
<td>clEnqueueWriteBuffer</td>
<td>8</td>
<td>0.109</td>
<td>0.013625</td>
<td>0.039</td>
<td>0.009</td>
</tr>
<tr>
<td>clEnqueueReadBuffer</td>
<td>8</td>
<td>0.083</td>
<td>0.010375</td>
<td>0.014</td>
<td>0.009</td>
</tr>
<tr>
<td>clEnqueueBuffer</td>
<td>8</td>
<td>0.054</td>
<td>0.00675</td>
<td>0.016</td>
<td>0.002</td>
</tr>
<tr>
<td>clEnqueueNDRangeKernel</td>
<td>2</td>
<td>0.048</td>
<td>0.024</td>
<td>0.025</td>
<td>0.023</td>
</tr>
<tr>
<td>clCreateKernel</td>
<td>2</td>
<td>0.031</td>
<td>0.0155</td>
<td>0.022</td>
<td>0.009</td>
</tr>
<tr>
<td>clSetKernelArg</td>
<td>8</td>
<td>0.019</td>
<td>0.002375</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>clGetDeviceInfo</td>
<td>3</td>
<td>0.012</td>
<td>0.004</td>
<td>0.008</td>
<td>0.002</td>
</tr>
<tr>
<td>clGetPlatformInfo</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>clBuildProgram</td>
<td>1</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>clCreateCommandQueue</td>
<td>1</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>clGetPlatformIDs</td>
<td>2</td>
<td>0.006</td>
<td>0.003</td>
<td>0.005</td>
<td>0.001</td>
</tr>
<tr>
<td>clCreateContextFromType</td>
<td>1</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>clGetDeviceIDs</td>
<td>1</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Kernel Execution Summary

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>Number Of Calls</th>
<th>Total Time (ms)</th>
<th>Average Time (ms)</th>
<th>Maximum Time (ms)</th>
<th>Minimum Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>smithwaterman</td>
<td>2</td>
<td>0.812</td>
<td>0.406</td>
<td>0.786</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Figure 9-1: SDAccel Profile Summary - Part 1
Chapter 9: Host Application Profiling in SDAccel

Kernel Execution Summary

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>Number Of Calls</th>
<th>Total Time (ms)</th>
<th>Average Time (ms)</th>
<th>Maximum Time (ms)</th>
<th>Minimum Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>smashwaterman</td>
<td>2</td>
<td>0.812</td>
<td>0.406</td>
<td>0.786</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Data Transfer Summary

<table>
<thead>
<tr>
<th>Transfer Type</th>
<th>Number Of Calls</th>
<th>Total Time (ms)</th>
<th>Average Time (ms)</th>
<th>Transfer Rate (MB/s)</th>
<th>Average Size (KB)</th>
<th>Maximum Size (KB)</th>
<th>Minimum Size (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOST READ BUFFER</td>
<td>8</td>
<td>0.014</td>
<td>0.00175</td>
<td>295.375</td>
<td>0.586</td>
<td>1.156</td>
<td>0.017</td>
</tr>
<tr>
<td>HOST WRITE BUFFER</td>
<td>8</td>
<td>0.017</td>
<td>0.002125</td>
<td>269.167</td>
<td>0.586</td>
<td>1.156</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Top Ten Kernel Execution Summary

<table>
<thead>
<tr>
<th>Kernel Name</th>
<th>Context ID</th>
<th>Command Queue ID</th>
<th>Device Name</th>
<th>Time (ms)</th>
<th>Global Work Size</th>
<th>Work Group Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>smashwaterman</td>
<td>1081429544</td>
<td>0</td>
<td>fpga0</td>
<td>0.786</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>smashwaterman</td>
<td>1081429544</td>
<td>0</td>
<td>fpga0</td>
<td>0.026</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Top Ten Buffer Writes

<table>
<thead>
<tr>
<th>Context ID</th>
<th>Command Queue ID</th>
<th>Time (ms)</th>
<th>Buffer Size (KB)</th>
<th>Writing Rate (MB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.003</td>
<td>1.156</td>
<td>385.333</td>
</tr>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.002</td>
<td>0.017</td>
<td>8.5</td>
</tr>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.002</td>
<td>1.156</td>
<td>8.5</td>
</tr>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.002</td>
<td>0.017</td>
<td>8.5</td>
</tr>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.002</td>
<td>1.156</td>
<td>8.5</td>
</tr>
<tr>
<td>1081429544</td>
<td>0</td>
<td>0.002</td>
<td>0.017</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Figure 9-2: SDAccel Profile Summary - Part 2

The profile summary includes a number of useful statistics for your OpenCL™ application. This can provide you with a general idea of the functional bottlenecks in your application. The profile summary consists of six sections.

API Calls Summary

This section displays the profile data for all OpenCL host API function calls executed in the host application.
Kernel Execution Summary

This section displays the profile data for all kernel functions scheduled and executed by the host application.

Data Transfer Summary

This section displays the profile data for all read and write transfers between the host and device memory via PCIe® link.

Top Ten Kernel Execution Summary

This section displays the profile data for top 10 kernels in terms of execution time. The table is ordered from the longest to shortest execution time.

Top Ten Buffer Writes

This section displays the profile data for top 10 write transfers from the host to the device memory in terms of execution time. The table is ordered from the longest to shortest execution time.

Top Ten Buffer Reads

This section displays the profile data for top 10 read transfers from the device memory to the host memory in terms of execution time. The table is ordered from the longest to shortest execution time.
Chapter 10

Xilinx OpenCL Compiler (xocc)

The Xilinx® OpenCL™ Compiler (xocc) is a standalone command line utility for compiling an OpenCL kernel supporting all flows in the SDAccel environment. It provides a mechanism for command line users to compile their kernels without writing any Tcl script, which is ideal for compiling host applications and kernels using a makefile.

xocc also supports clCreateProgramWithSource() OpenCL API. The SDAccel runtime library automatically invokes xocc when this function is called from a host application.

Following are details of xocc command line format and options.

Syntax:

```
xocc [options] <input_file>
```

<table>
<thead>
<tr>
<th>Options</th>
<th>Valid Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>--xdevice &lt;arg&gt;</td>
<td>Supported acceleration devices by Xilinx and third-party board partners.</td>
<td>Set target Xilinx device. This option is mandatory. The devices below are installed with the SDAccel installation: xilinx:adm-pcie-7v3:1ddr:1.0 : ADM-PCIE-7V3 Card xilinx:pico_m505:pico:1.0 : Pico MX505-EX400 Card</td>
</tr>
<tr>
<td>-t [sw_emu</td>
<td>hw_emu</td>
<td>hw]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• sw_emu: CPU emulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• hw_emu: hardware emulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• hw: hardware</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Default: hw</td>
</tr>
<tr>
<td>-o &lt;arg&gt;</td>
<td>Any valid file name with .xclbin extension</td>
<td>Set output binary file name.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Default: a.xclbin</td>
</tr>
<tr>
<td>-D &lt;name=definition&gt;</td>
<td></td>
<td>Predefine name as a macro with definition. This option is passed to the openCL preprocessor.</td>
</tr>
</tbody>
</table>
Following is a makefile example showing how to compile host applications outside the SDAccel environment, and use xocc to compile kernels for CPU and hardware emulation as well as targeting hardware.

**IMPORTANT:** `g++` from SDAccel installation directory is required to compile host applications outside SDAccel environment.

```makefile
<SDAccel_installation_path>/lnx64/tools/gcc/bin/g++
```

<table>
<thead>
<tr>
<th>Options</th>
<th>Valid Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-I &lt;directory_name&gt;</td>
<td></td>
<td>Add the directory to the list of directories to be searched for header files. This option is passed to the openCL preprocessor.</td>
</tr>
<tr>
<td>-h</td>
<td>NA</td>
<td>Print help</td>
</tr>
<tr>
<td>-s</td>
<td>NA</td>
<td>Do not delete intermediate files</td>
</tr>
<tr>
<td>-g</td>
<td>NA</td>
<td>Generate code for debugging</td>
</tr>
<tr>
<td>input file</td>
<td>OpenCL kernel file name with .cl extension</td>
<td>All kernels need to be put into a single kernel file and they will be compiled into the same XCLBIN file.</td>
</tr>
</tbody>
</table>

Table 10-1: xocc Options

Following is a makefile example showing how to compile host applications outside the SDAccel environment, and use xocc to compile kernels for CPU and hardware emulation as well as targeting hardware.

**IMPORTANT:** `g++` from SDAccel installation directory is required to compile host applications outside SDAccel environment.

```makefile
<SDAccel_installation_path>/lnx64/tools/gcc/bin/g++
```

```makefile
VPATH = ./

#supported flow: cpu_emu, hw_emu, hw
FLOW=hw
HOST_SRCS = main.c
HOST_EXE = hello_world
HOST_CFLAGS = -DFPGA_DEVICE -g -Wall -I${XILINX_OPENCL}/runtime/include/1_2
HOST_LFLAGS = -L${XILINX_OPENCL}/runtime/lib/x86_64 -lxilinxopencl

KERNEL_SRCS = hello_world.cl
KERNEL_DEFS =
KERNEL_INCS =

#set target device for XCLBIN
#Set target Xilinx device. Devices below are installed with SDAccel installation:
# xilinx:adm-pcie-7v3:1ddr:1.0 : ADM-PCIE-7V3 Card
# xilinx:pico_m505:pico:1.0 : Pico MX505-EX400 Card
#XDEVICE=xilinx:adm-pcie-7v3:1ddr:1.0
#KEEP_TEMP=
KERNEL_DEBUG=

CC = g++
CLCC = xocc
CLCC_OPT = --device ${XDEVICE} -o ${XCLBIN} ${KERNEL_DEFS} ${KERNEL_INCS}

ifeq ($(FLOW),cpu_emu)
    CLCC_OPT += -t sw_emu
    XCLBIN = hello_world_cpu_emu.xclbin
else ifeq ($(FLOW),hw_emu)
    CLCC_OPT += -t hw_emu
    XCLBIN = hello_world_hw_emu.xclbin
else ifeq ($(FLOW),hw)
    CLCC_OPT += -t hw
    XCLBIN = hello_world_hw.xclbin
    CLCC_OPT += -t hw
```
endif

ifeq (${KEEP_TEMP},1)
    CLCC_OPT += -s
endif

ifeq (${KERNEL_DEBUG},1)
    CLCC_OPT += -g
endif

HOST_ARGS = ${XCLBIN}
OBJECTS := ${HOST_SRCS:.c=.o)

.PHONY: all

all:

run : host xbin
    ${HOST_EXE} $(HOST_ARGS)

host: $(HOST_EXE)

xbin : $(XCLBIN)

$(HOST_EXE): $(OBJECTS)
    $(CC) $(HOST_LFLAGS) $(OBJECTS) -o $@

$(XCLBIN): $(KERNEL_SRCS)
    $(CC) $(CLCC_OPT) $(KERNEL_SRCS)
%
.o: %.c
    $(CC) $(HOST_CFLAGS) -c $< -o $@

clean:
    $(RM) $(HOST_EXE) $(OBJECTS) $(XCLBIN)
Overview

SDAccel™ is a command line based environment for the development of OpenCL™ applications targeted at Xilinx® FPGA devices. This appendix provides a reference for every command in SDAccel.

Solution Creation Commands

create_solution

Syntax

    create_solution -name <solution name> -dir <solution directory> -force

Description

SDAccel is a solution based tool. All applications compiled by SDAccel must be part of a solution. The create_solution command defines a solution within the SDAccel development environment.

Options

-name

The name option defines the name of the solution in the SDAccel environment. The solution name can be any name defined by the user. It is used to create the top level working directory where the application will be compiled.

-dir

The dir option allows the user to specify the directory where the SDAccel solution is to be created. This setting is optional. By default, SDAccel creates the solution in the current directory from where the tool was invoked.
**-force**

The `force` option tells SDAccel to completely delete and overwrite the solution directory specified by the name option if it already exists.

**Examples**

Create a new solution name `mysolution` in the temp directory and delete any previous work for `mysolution` if it exists.

```bash
create_solution -name mysolution -dir /temp -force
```

---

## Add Target Device

**Syntax**

```bash
add_device <device_name>
```

**Description**

The `add_device` command defines the target device for which the application is compiled. The name of the device is determined by the device provider for SDAccel. A list of devices included in SDAccel is available in Appendix B, Devices.

**Example**

```bash
add_device xilinx:adm-pcie-7v3:1ddr:1.0
```

---

## Host Code Commands

**add_files**

**Syntax**

```bash
add_files <file name>
```
Description

The `add_files` command adds user code host application files to the current solution in SDAccel. The files listed by the `add_files` command are used from their current location in the programmer's environment. Files are not copied nor stored in the SDAccel solution.

Example

```plaintext
add_files foo.c
add_files foo.h
```

---

Host Code Properties

**file_type**

Syntax

```plaintext
set_property file_type "c header files" [get_files <file name>]
```

Description

The `add_files` command handles both header and source files in the same way. For proper compilation of the user application, SDAccel needs to know which files to treat as header files during the code compilation phase. The `file_type` property allows the user to specify which files to treat as header files. The `get_files` command that is attached to this property defines which file in the solution the `file_type` property is used with.

Example

```plaintext
set_property file_type "c header files" [get_files "foo.h"]
```

**host_cflags**

Syntax

```plaintext
set_property -name host_cflags -value "<compiler flags>" -objects [current_solution]
```

Description

The `host_cflags` property defines any additional compiler flags that must be set during the compilation of the host code. These flags are passed as is by SDAccel to the code compiler. The code compiler used by SDAccel to process the host code depends on the design phase of the application and the host processor type in the target device.

Example

```plaintext
set_property -name host_cflags -value "-g -D XILINX" -objects [current_solution]
```
Kernel Definition Commands

create_kernel

Syntax

    create_kernel <kernel name> -type <kernel source code type>

Description

The create_kernel command defines a kernel within the context of the current solution. The name of the kernel must match the name of the kernel function in the source code whose operations will be compiled into a custom compute unit for execution on the FPGA fabric.

Options

-type

The type option defines the language in which the kernel has been captured. The valid types in SDAccel are clc for kernels expressed in OpenCL C kernel language and c for kernels expressed in C/C++ that can be consumed by the Vivado® HLS tool.

Example

    create_kernel test -type clc

Kernel Code Commands

add_files

Syntax

    add_files -kernel [get_kernels <kernel name>] <kernel source file name>

Description

The add_files command with the kernel option adds user kernel code files to the current solution in SDAccel. Since SDAccel supports more than one kernel per solution, the add_files command also has to specify which kernel in the solution do the source files correspond to. The correspondence between source files and kernels in the solution is achieved by the get_kernels command.
Example

```
add_files -kernel [get_kernels test] test.cl
```

---

## Kernel Properties

### kernel_flags

**Syntax**

```
set_property kernel_flags <compiler flags> [get_kernels <kernel name>]
```

**Description**

The `kernel_flags` property defines any additional compiler flags that must be set during the compilation of kernel code. Kernel flags are set on a per kernel basis and each kernel can be compiled with different flags. The pairing between compiler flags and the kernel object in the SDAccel solution is through the use of the `get_kernels` command.

**Example**

```
set_property kernel_flags "-D" [get_kernels test]
```

### max_memory_ports

**Syntax**

```
set_property max_memory_ports true [get_kernels <kernel name>]
```

**Description**

The `max_memory_ports` property is used for memory architecture optimization. This property informs SDAccel to create one physical memory port per global memory buffer defined in the kernel code. This property applies to all compute units generated for a given kernel.

**Example**

```
set_property max_memory_ports true [get_kernels test]
```

### memory_data_port_width

**Syntax**

```
set_property memory_data_port_width <port width> [get_kernels <kernel name>]
```

Appendix A: Command Reference

Description

The `memory_data_port_width` property is used for memory architecture optimization. This property informs SDAccel to override the fundamental bitwidth of a buffer data type and implement an interface of size port width. The supported interface sizes in SDAccel for memory interfaces are 32, 64, 128, 256, and 512 bits. The setting of this property applies to all compute units generated for a given kernel.

Example

```
set_property memory_data_port_width 512 [get_kernels test]
```

Binary Container Definition Commands

create_opencl_binary

Syntax

```
create_opencl_binary -device <device name> <binary name>
```

Description

SDAccel creates precompiled binaries that contain the custom compute units, which will be executed concurrently on the FPGA fabric. A binary is compiled to run on a specific device of the device. The name of the binary can be anything the user wants.

Options

- `device`

The `device` property defines which device in the target device will be used to execute the OpenCL binary container.

Example

```
create_opencl_binary -device "xilinx:adm-pcie-7v3:1ddr:1.0" test_binary
```

Binary Container Properties

region

Syntax

```
set_property region <region name> [get_opencl_binary <binary name>]
```
Description

The `region` property is an optional property that helps further define where on a device the current binary container is executed. All devices supported by SDAccel have at least one region. By default, SDAccel chooses the first computation region available for the device selected during the definition of the binary container.

Example

```
set_property region "OCL_REGION_0" [get_opencl_binary test_binary]
```

---

## Compute Unit Definition Commands

### create_compute_unit

**Syntax**

```
create_compute_unit -opencl_binary [get_opencl_binary <binary name>]
               -kernel [get_kernels <kernel name>]
               -name <compute unit instance name>
```

**Description**

The `create_compute_unit` command creates an instance of the custom logic required to execute the operations of a given kernel. Compute units are created within the context of a binary container. A binary container can contain multiple compute units for the same kernel.

**Options**

- `-opencl_binary`
  
The `opencl_binary` option defines which binary container the current compute unit is to be created in.

- `-kernel`
  
The `kernel` option defines the kernel in the solution from which the compute unit is created.

- `-name`
  
The `name` tag is an optional command to distinguish between compute units in the system. If a kernel only has one compute unit, the name tag is optional. If a kernel has multiple compute units in the system, each compute unit must have a unique name. The name of the compute unit is used internally by the Xilinx OpenCL runtime library to schedule the execution of kernel operations. This name is not exposed to the application programmer in any way and is not referred to in the application host code.
Example

create_compute_unit -opencl_binary [get_opencl_binary test_binary]
-kernel [get_kernels test]
-name myinstance

---

Emulation Commands

`compile_emulation`

**Syntax**

```
compile_emulation -flow {cpu | hardware}
```

**Description**

The `compile_emulation` command compiles the current solution for execution in the emulation environment provided by SDAccel. There are two emulation modes supported in SDAccel. These are CPU and hardware. The CPU emulation is the default emulation environment and allows the programmer a fast turnaround for compile, debug, and run iterations while the application functionality is being developed. The hardware emulation flow executes the custom logic generated for each compute unit inside of a hardware simulator to provide the application programmer with assurance that the compiler built the correct logic before testing it on the FPGA.

**Options**

- **-flow**

  The `flow` option determines if the application is to be compiled for CPU or hardware emulation. CPU emulation is the default. Therefore this flag is only required when selecting compilation for the hardware emulation flow.

**Example**

- CPU Emulation
  
  `compile_emulation`

- Hardware Emulation
  
  `compile_emulation -flow hardware`

`run_emulation`

**Syntax**

```
run_emulation -flow {cpu | hardware} -args <command line arguments>
```
Description

The run_emulation command executes the compiled application in either the CPU or hardware emulation environments. The arguments flag provides a mechanism to provide command line arguments to the host program executable. The run_emulation command must be preceded by a compile_emulation command.

Options

-flow

The flow options allows the user to determine if the application will be executed in either the CPU or the hardware emulation environments. The default is to execute all applications using CPU based emulation. Therefore, this option is only needed when specifying that an application must be executed in the hardware emulation environment.

-args

The args option enables the passing of command line arguments to the host program executable. Any option that would be entered at the command line with the host program executable can be used here.

Example

- CPU Emulation
  
  run_emulation

- Hardware Emulation
  
  run_emulation -flow hardware

Reporting Commands

report_estimate

Syntax

  report_estimate
Appendix A: Command Reference

Description

The `report_estimate` command generates a compute unit level estimate on the application performance the FPGA resources required to implement the functionality. It is a good practice to run this command and analyze its results before leaving the emulation part of the design cycle and moving to the FPGA system compilation step.

Example

```bash
report_estimate
```
Description

The `build_system` command creates the binaries for the host code and the compute units to be executed on FPGA based accelerator cards. The `package_system` command completes the binary creation process by placing all generated binaries in a central location for the application programmer to use for deployment. All binaries suitable for execution on the FPGA accelerator card are placed in this location:

```
<solution name>/pkg
```

Example

```
package_system
```

---

**System Execution Commands**

**run_system**

**Syntax**

```
run_system -args <command_line_arguments>
```

**Description**

The `run_system` command allows the user to execute the application on the FPGA based accelerator card from within the SDAccel environment. This is identical to taking the compiled binaries in the `pkg` directory and executing them in a shell outside of SDAccel.

**Example**

```
run_system -args "test.xclbin"
```
Appendix B

Devices

SDAccel Supported Devices

SDAccel™ solutions are compiled against a target device. A device is the combination of board and infrastructure components on which the kernels of an application is executed. Applications compiled against a target device can only be executed on the board associated with the target device. Devices can be provided by SDAccel ecosystem partners, FPGA design teams, and Xilinx®. Contact your Xilinx representative for more information about adding third party devices in SDAccel. The following table shows the devices included in SDAccel:

Table B-1:  SDAccel Supported Devices

<table>
<thead>
<tr>
<th>Device</th>
<th>Board</th>
<th>FPGA Device</th>
<th>Board Vendor</th>
</tr>
</thead>
<tbody>
<tr>
<td>xilinx:adm-pcie-7v3:1ddr:1.0</td>
<td>ADM-PCIE-7V3</td>
<td>Virtex-7 690T</td>
<td>Alpha Data</td>
</tr>
<tr>
<td>xilinx:pico_m505:pico:1.0</td>
<td>M-505</td>
<td>Kintex-7 325T</td>
<td>Pico Computing</td>
</tr>
</tbody>
</table>


Appendix C

Additional Resources and Legal Notices

Xilinx Resources

For support resources such as Answers, Documentation, Downloads, and Forums, see Xilinx Support.

For a glossary of technical terms used in Xilinx documentation, see the Xilinx Glossary.

Solution Centers

See the Xilinx Solution Centers for support on devices, software tools, and intellectual property at all stages of the design cycle. Topics include design assistance, advisories, and troubleshooting tips.

References

2. SDAccel Development Environment Installation and Licensing Guide, (UG1020)

Please Read: Important Legal Notices

The information disclosed to you hereunder (the “Materials”) is provided solely for the selection and use of Xilinx products. To the maximum extent permitted by applicable law: (1) Materials are made available “AS IS” and with all faults, Xilinx hereby DISCLAIMS ALL WARRANTIES AND CONDITIONS, EXPRESS, IMPLIED, OR STATUTORY, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE; and (2) Xilinx shall not be liable (whether in contract or tort, including negligence, or under any other theory of liability) for any loss or damage of any kind or nature related to, arising under, or in connection with, the Materials (including your use of the Materials), including for any direct, indirect, special, incidental, or consequential loss or damage (including loss of data, profits, goodwill, or any type of loss or damage suffered as a result of any action brought by a third party) even if such damage or loss was reasonably foreseeable or Xilinx had been advised of the possibility of the same. Xilinx assumes no obligation to correct any errors contained in the Materials or to notify you of updates to the Materials or to product specifications. You may not reproduce, modify, distribute, or publicly display the Materials without prior written consent. Certain products are subject to the terms and conditions of Xilinx’s limited warranty, please refer to Xilinx’s Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos; IP cores may be subject to warranty and support terms contained in a license issued to you by Xilinx. Xilinx products are not designed or intended to be
fail-safe or for use in any application requiring fail-safe performance; you assume sole risk and liability for use of Xilinx products in such critical applications, please refer to Xilinx's Terms of Sale which can be viewed at http://www.xilinx.com/legal.htm#tos.

NOTICE: This pre-release document contains confidential and proprietary information of Xilinx, Inc. and is being disclosed to you as a participant in an early access program, under obligation of confidentiality. You may print one (1) copy of this document for evaluation purposes. You may not modify, distribute, or disclose this material to anyone, including your employees, coworkers, or contractors (these individuals must apply for separate access to the program). This document contains preliminary information and is subject to change without notice. Information provided herein relates to products and/or services not yet available for sale, and provided solely for information purposes and are not intended, or to be construed, as an offer for sale or an attempted commercialization of the products and/or services referred to herein.

© Copyright 2015 Xilinx, Inc. Xilinx, the Xilinx logo, Artix, ISE, Kintex, Spartan, Virtex, Vivado, Zynq, and other designated brands included herein are trademarks of Xilinx in the United States and other countries. All other trademarks are the property of their respective owners.